Summer 8-15-2013

The Enhancing Effect of Retrieval on Subsequent Encoding: Understanding Test-Potentiated Learning

Kathleen Marie Arnold
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychology

Dissertation Examination Committee:
Kathleen B. McDermott, Chair
David A. Balota
Joe Barcroft
Susan M. Fitzpatrick
Mark A. McDaniel
Henry L. Roediger, III

The Enhancing Effect of Retrieval on Subsequent Encoding: Understanding Test-Potentiated Learning

by

Kathleen Marie Arnold

A dissertation presented to the
Graduate School of the Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2013

St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

List of Appendices

Acknowledgments

I want to offer my sincere gratitude to the numerous people without whom this dissertation would never have been completed. Throughout my education there have been numerous mentors, teachers, colleagues, students, friends, and family who have taught, guided, and supported me. Only with their help have I been able to achieve this great milestone.

First, I will be forever thankful to my graduate school advisor and chair of my dissertation committee, Kathleen McDermott. Throughout my six years at Washington University in St. Louis, she has supported and guided me as I faced the challenges of graduate school. From her I have learned how to be a more careful scientist, critical thinker, and proficient communicator. I am deeply appreciative of the scientific foundation that has she has provided me, and as I continue in my career and build upon that foundation, I look forward to continuing our relationship as it advances from teacher-student to colleagues and friends.

I would also like to thank Roddy Roediger who has been somewhat of a second mentor throughout my graduate career. I am grateful for his generosity in sharing his wisdom and passion with me. Having him as an unofficial second advisor has added immeasurable value to my education. I am also thankful to the other members of my committee, Dave Balota, Mark McDaniel, Joe Barcroft, and Susan Fitzpatrick, who provided invaluable input and guidance throughout the dissertation process. I thank them for donating their time and expertise to help me improve my project and become a better scientist.

Next, I would like to thank several people who were instrumental in helping me complete this dissertation. Andrew Fishell, Fan Zou, Ana Hernandez, and Sophie Crumpacker all provided valuable assistance in helping me program the experiment and/or score and code the data. I would like to acknowledge Andrew's contribution in particular for his incredibly helpful

assistance in programming the experiments in Flash. I thank Mike Strube and Josh Jackson for acting as statistical consultants and providing helpful discussions about my data analyses. I also want to acknowledge the many helpful discussions I had with members of the Memory and Cognition lab and the Memory lab that provided useful advice and suggestions to improve the project.

In addition to providing help and support throughout the dissertation process, the members of the Memory and Cognition lab have been a vital part of my entire graduate career. I want to thank former and current members of the Memory and Cognition Lab who have been my compatriots throughout the years: Karl Szpunar, Adrian Gilmore, Sean Kang, Hank Chen, Cindy Fadler, Bridgid Finn, Steve Nelson, Mary Pyc, and Farah Naaz. They have intellectually challenged me, emotionally supported me, and materially aided me in many different ways. I also want to thank my many colleagues from other labs who have helped me manage the challenges of graduate school. The members of my cohort, with whom I tackled the many milestones of our program, – especially Emily Bloesch, Alexandra Zaleta, Geoff Maddox, and Kim Chiew – and the many other friends from the surrounding cohorts helped make graduate school a wonderful – albeit challenging, stressful, and, at times, frustrating - experience.

I would be remiss if I did not also acknowledge my undergraduate mentor, Gil Einstein. Without him, I would never have entered graduate school or pursued cognitive psychology. With his passionate teaching and generous mentorship, he instilled in me what I assume will be a life-long love of psychology.

Finally, I want to thank my family, especially my parents. Without their support, encouragement, and love I am sure that I would not be getting my Ph.D. today. I want to thank them for always emphasizing education and teaching me the value of learning. They have always

supported me in whatever I have chosen to do and that has given me the strength to pursue my passion. I also want to thank my brothers – Joe, Chris, and Matt – for being my compatriots through life. Even though growing up there were times they aggravated me, annoyed me, and made me cry (as brothers do), throughout the years they have also toughened me up, supported me, made me laugh, and provided me with joy and pride. I am happy to call them not just brothers but friends. I also want to thank my extended family. They provided a loving environment for me to grow up in and have continued to support and encourage me.

This research was supported by a Collaborative Activity Grant from the James S. McDonnell Foundation and a Dissertation Fellowship from Washington University in St. Louis.

ABSTRACT OF THE DISSERTATION

The Enhancing Effect of Retrieval on Subsequent Encoding: Understanding Test-Potentiated Learning

by

Kathleen Marie Arnold

Doctor of Philosophy in Psychology

Washington University in St. Louis, 2013

Professor Kathleen McDermott, Chair

Retrieval practice directly enhances later memory of tested material, a robust effect known as the *testing effect* (Roediger & Karpicke, 2006b). Numerous experiments have provided support for this effect. However, another important effect of retrieval practice has received far less attention. Retrieval practice can also indirectly enhance learning by potentiating subsequent encoding of tested material, an effect known as *test-potentiated learning* (Izawa, 1966). Although introduced over four decades ago, little is known about how and when tests enhance subsequent encoding, information that has both practical and theoretical importance. The aim of this dissertation was to enhance understanding of test-potentiated learning by answering two fundamental questions: First, which aspect(s) of prior tests (unsuccessful retrieval, successful retrieval, and/or spacing) cause subsequent enhanced encoding (Experiment 1) and second, can items that are correctly retrieved prior to restudying benefit from test-potentiated learning (Experiment 2)?

Experiment 1 answered the first question by varying the amount of unsuccessful and successful retrieval subjects engaged in prior to restudying the material. Further, the lag between study periods, task engagement, and exposure to recallable items were held constant across

conditions to equate the effect of spacing. Repeated unsuccessful retrieval enhanced the benefit of the restudy trial for initially incorrect items. The implications of this finding for the theoretical processe(s) driving test-potentiated learning are discussed.

Experiment 2 tested the hypothesis that items correctly retrieved prior to restudying can also benefit from test-potentiated learning as evidenced by later enhanced recall. A restudy opportunity has been shown to enhance later recall of previously retrieved items, especially items retrieved with low confidence (Butler, Karpicke, & Roediger, 2008). That is, restudying enhances the effect of prior retrieval. Prior retrieval may also enhance the effect of subsequent restudy. This hypothesis was tested by varying the number of prior tests and whether or not subjects had a restudy opportunity. Further, confidence ratings were collected. Restudying was found to benefit low-confident correct items, but the number of prior tests did not modify this effect suggesting that correct retrieval prior to restudying does not enhance the effect of restudying on later recall. Together, these experiments indicate that unsuccessful retrieval attempts made during prior testing are the driving force behind test-potentiated learning.

Retrieval is a process of memory not a static outcome, and as such it has the capacity to dynamically affect memory. Some of these dynamic effects have been studied in depth such as the effect of retrieval on later memory. For instance, consider a learner attempting to master a large set of vocabulary words. Practicing retrieval of the definitions (i.e., attempting to recall the definitions such as by using flashcards) will enhance the likelihood that the learner will be able to recall those definitions on a later test. That is, retrieving an item from memory enhances the probability that the item will be retrieved in the future, a robust effect commonly referred to as the *testing effect* (for reviews, see McDermott, Arnold, & Nelson, in press; Roediger & Butler, 2011; Roediger & Karpicke, 2006b).

The testing effect is a direct effect of retrieval in that the act of retrieving directly affects later recall. Retrieval can also indirectly affect memory. One such indirect effect is the enhancing effect of retrieval on subsequent encoding opportunities. If the learner attempting to master the vocabulary words attempts to retrieve the definitions prior to restudying, learning will be enhanced relative to if the learner had restudied the definitions without first engaging in retrieval practice. That is, prior retrieval practice enhances subsequent encoding, an effect known as *test-potentiated learning* (Izawa, 1971). This effect has received far less attention than the testing effect and therefore is not as well understood, which is particularly unfortunate because of its potential importance in educational settings.

In this dissertation, test-potentiated learning is defined as the enhanced learning that takes place during a restudy opportunity when one or more tests have been taken between an initial study and restudy period. Importantly, the intervening test(s) must involve episodic retrieval (or attempted retrieval) of the previously studied material, and the restudy opportunity must involve restudy of that same material. Throughout this dissertation, the first intervening test will be

referred to as the initial test. A later test taken after the restudy opportunity will be referred to as the final test. The final test is given to measure the amount of learning that has taken place during restudy.

Test-potentiated learning results in greater recall on a final test in a condition in which one or more tests preceded a restudy opportunity as compared to a condition in which no or fewer tests preceded a restudy opportunity. However, given that tests can have both direct and indirect effects on later recall, greater recall on a final test in the former condition does not necessarily indicate that test-potentiated learning occurred. A way to distinguish the effects of test-potentiated learning from the testing effect is needed to measure the manifestation of test-potentiated learning. One way this can be done is by including control conditions in which there is no restudy opportunity, which allows the effect of restudy to be measured by comparing final recall in conditions with and without a restudy opportunity. The presence of test-potentiated learning is inferred when the effect of restudying is larger in conditions with more intervening tests. For example, Arnold and McDermott (2013) used a 2 (number of intervening tests: 0, 3) X 2 (restudy after intervening tests, no restudy after intervening tests) factorial design to measure the effect of test-potentiated learning. As can be seen in Figure 1, the benefit of having a restudy opportunity was larger when more intervening tests had been taken indicating the presence of test-potentiated learning.

There are two types of items that could benefit from test-potentiated learning: initially incorrect items (items not retrieved on the initial test) and initially correct items (items

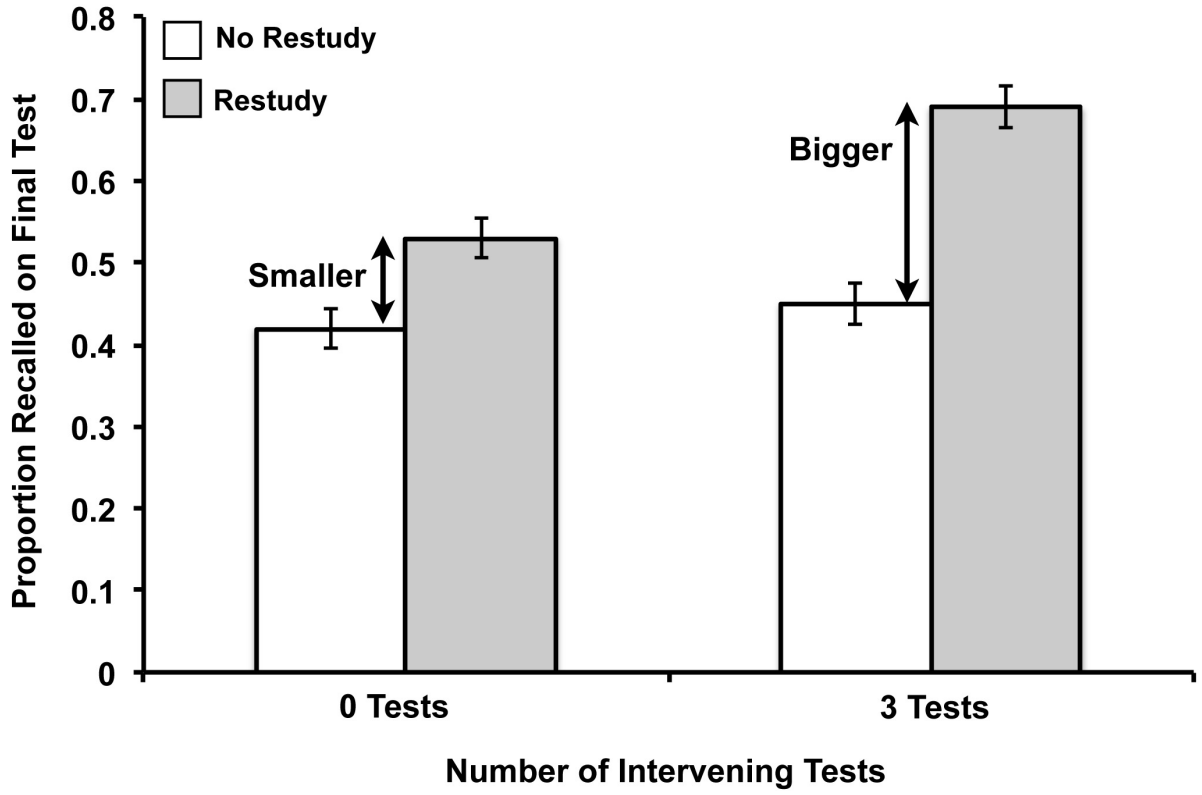

Figure 1. Mean proportion of items recalled on the final test in Experiment 1 of Arnold and McDermott (2013) as a function of the number of initial tests and whether or not the items were restudied. Error bars represent standard errors of the means. Adapted from their Figure 2.

successfully retrieved on the initial test)[1]. First, subsequent learning of items that had been attempted, but failed, to be retrieved on an initial test could be potentiated. This is the traditional conceptualization of test-potentiated learning and will be referred to as the *enhanced encoding* component. Several prior studies have provided supporting evidence for enhanced encoding (e.g., Arnold & McDermott, 2012; in preparation; Izawa, 1968, 1971).

Second, retrieval practice prior to restudying could enhance the effect of the restudy opportunity on retention, or the probability that items recalled on the initial test are also recalled on the final test. Throughout this dissertation, retention will be measured as the proportion of initially correctly retrieved items that are also retrieved on the final test. Previous research has shown that providing feedback (i.e., a restudy opportunity) after successful retrieval can enhance retention relative to a no feedback condition, especially when items have been initially retrieved with low confidence (Butler, Karpicke, & Roediger, 2008). That is, feedback can modify the effect of successful retrieval. The converse may also be true; successful retrieval may modify the effect of feedback or restudy. Successfully retrieving an item prior to restudying could enhance subsequent encoding of that item. That is, a restudy opportunity following a test may enhance retention more than a restudy opportunity that does not follow a test. This effect will be referred to as the *enhanced retention* component of test-potentiated learning. Because this is not the traditional conceptualization of test-potentiated learning, no prior studies have found direct evidence supporting this component.

[1] A third type of item could also potentially benefit from test-potentiated learning: items that are not initially tested. That is, encoding on a restudy opportunity of items that had been initially studied but had not been initially tested could be enhanced by prior retrieval practice of other items. However, previous research suggests that these items are not potentiated (Arnold, Nelson, & McDermott, in preparation), and they are outside of the scope of the current experiments (see General Discussion).

The primary aim of the present research is to better understand and conceptualize test-potentiated learning. To do this, both components of test-potentiated learning, enhanced encoding and enhanced retention, are examined separately. Experiment 1 focuses on enhanced encoding. Although previous experiments have provided support for the existence of this effect, there has been little work on understanding why tests potentiate learning of initially incorrect items. The first step in understanding why tests enhance subsequent encoding is to understand which aspect(s) of tests can enhance encoding. A typical test involves a mixture of unsuccessful and successful retrieval, either or both of which may enhance subsequent encoding. For instance, unsuccessful retrieval may lead to activation of partial knowledge (Grimaldi & Karpicke, 2012), and/or enhance participants' metacognitive knowledge (Arnold & McDermott, 2012; Lachman & Laughery, 1968). Alternatively, successful retrieval may lead to enhanced organizational processes (Arnold & McDermott, 2013) and/or reduced interference (Szpunar, McDermott, & Roediger, 2008). Further, tests take time and therefore produce a lag between study trials. Previous literature has shown that spacing between study trials can itself enhance subsequent learning (for a review, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006), and therefore tests may enhance subsequent encoding through spacing rather than through retrieval processes. Experiment 1 compares these three possible sources of enhanced encoding (unsuccessful retrieval, successful retrieval, and spacing) to determine which aspect(s) of tests potentiate learning of initially incorrect items.

Experiment 2 examines the enhanced retention component of test-potentiated learning. As previously mentioned, this component has not been traditionally included in the conceptualization of test-potentiated learning. For this reason, previous experiments have not specifically focused on this effect, and there is no direct evidence supporting this component.

This neglect is in part due to a common assumption that restudying or feedback has little or no benefit for items correctly recalled on an initial test (Anderson, Kulhavy, & Andre, 1971; Guthrie, 1971; Kulhavy & Anderson, 1972; Pashler, Cepeda, Wixted, & Rohrer, 2005; Pashler, Rohrer, Cepeda, & Carpenter, 2007). If restudying does not benefit initially correct items, then no amount of prior testing could enhance this nonexistent effect. However, Butler et al. (2008) found evidence that feedback does benefit initially correct items. Specifically, they found that later recall of low-confident correct items was enhanced when feedback was provided. Testing may enhance the effect of restudying, and this enhancement may be more pronounced for low-confident items. Experiment 2 tests this hypothesis by comparing the effect of restudying on retention in conditions with different amounts of prior retrieval practice.

In the following sections, I first review the relevant literature on test-potentiated learning with an emphasis on how previous experiments support (or fail to support) enhanced encoding and enhanced retention. I then briefly review the literature on three related effects: generate-potentiated learning, the testing effect, and the spacing effect. As will become clear, consideration of these effects aids in the understanding test-potentiated learning. Finally, I describe the current experiments in detail and discuss the implications of their results.

A historical overview of test-potentiated learning

The concept of test-potentiated learning was first introduced by Izawa (1966), who used a cued recall paradigm to investigate the role of tests in learning. Several studies were conducted following up on her findings and investigated the potentiating effects of cued recall tests. In a separate line of work, Tulving (1967) used a free recall paradigm to investigate the role of tests in learning and also found preliminary support for test-potentiated learning. Like Izawa's research, his study also spurred follow up experiments, but ones that investigated the potentiating

effects of free recall tests. Although these two lines of work are highly connected, they come from slightly different traditions. For this reason, each line of work is discussed separately. Past research on test-potentiated learning in cued recall paradigms is discussed first followed by a discussion on research on test-potentiated learning in free recall paradigms.

Cued recall test-potentiated learning

In the years after her first test-potentiated learning experiment (Izawa, 1966), Izawa (1967a, 1967b, 1968, 1969a, 1969b, 1970a, 1970b, 1971) replicated and extended her findings in multiple studies. In her experiments, she typically compared learning curves obtained from using different sequences of cued recall tests and study trials and generally found that learning on any given study trial was more efficient when more tests had been included in the sequence of previous trials. For example, in one experiment, Izawa (1971) compared five different learning sequences (see Figure 2). In between study trials, items were tested one to five times. When the results were analyzed as a function of the number of study trials, recall was higher in conditions with more intervening tests between study trials. For instance, as Figure 2 shows, after four study trials recall was highest in the five intervening test condition and lowest in the one intervening test condition. An alternative way to view the results is that when more intervening tests had been taken between study trials, fewer study trials were needed to reach the same level of recall. For instance, in the five intervening tests condition more than 68% of the items could be recalled after four study trials, whereas eight study trials were needed to reach this same level of recall in the one intervening test condition.

In addition to these results, Izawa (1966-1971) found that across consecutive cued recall tests there was no significant change in the proportion of words that were recalled. She

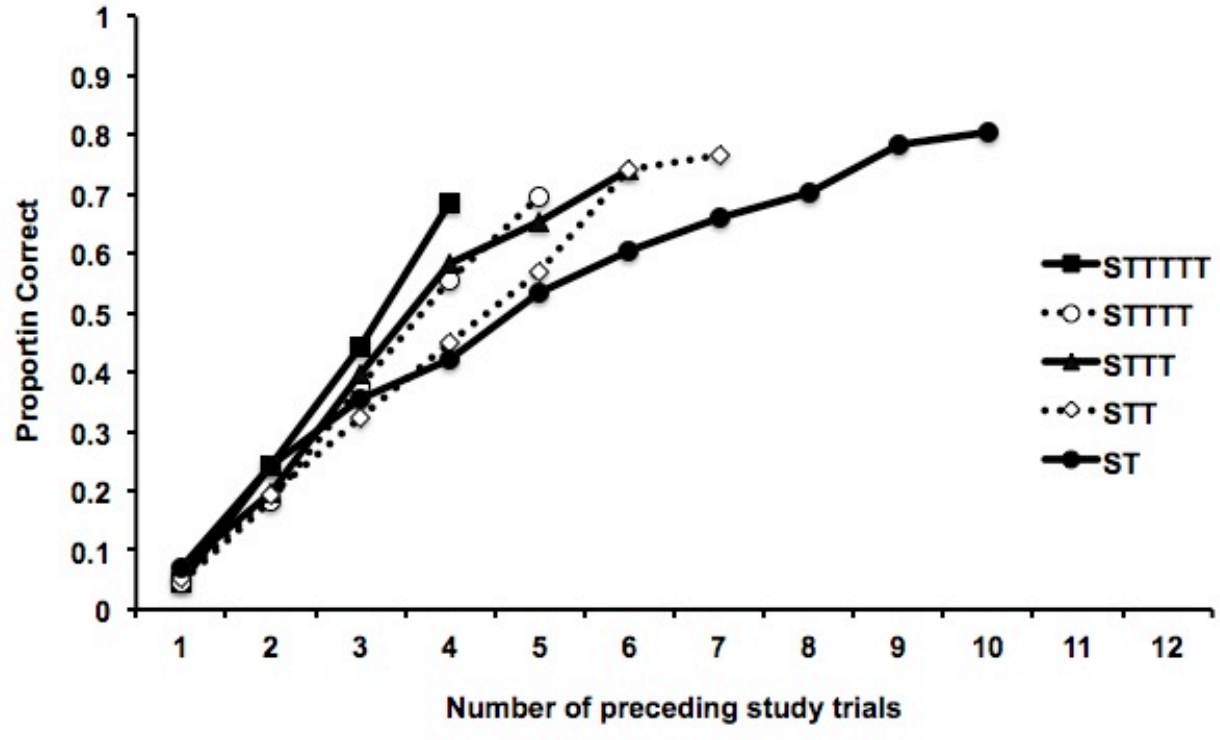

Figure 2. Mean proportion correct on the first test following each study trial in Experiment 1 of Izawa (1971) as a function of the number of preceding study trials. S = study trial, T = test trial. Data as reported in Table 2.

concluded that no learning or forgetting occurred during the test trials. Because of this finding, she posited that the observed recall differences between conditions had to be due to differences in learning during study trials. Therefore, she concluded that the tests were likely potentiating, or increasing the efficiency of, learning on subsequent study trials.

In many of Izawa's (1966, 1971) experiments, including the one described above, the number of test trials was confounded with spacing between study trials. Conditions with more test trials between study trials also had longer lags between study trials. Izawa was aware of this confound and did further experiments to determine if the enhancing effect was due to spacing rather than to testing. In one such follow-up study, Izawa (1968) conducted three experiments that compared learning in conditions with varying numbers of tests between study trials to learning in conditions with varying numbers of distractor task or blank trials between study trials. The distractor task involved naming geometric shapes and was unrelated the main task. On blank trials, participants were not given any task. On a final test, she found that given the same lag between study trials, recall was better in conditions with test trials rather than with distractor or blank trials between study trials. That is, testing between study trials enhanced learning relative to doing an unrelated or no task between study trials. From these results, Izawa concluded that tests have a special potentiating function that enhances subsequent learning above and beyond any enhancement gained from the lag between study trials.

Izawa (1966-1971) approached test-potentiated learning from a one-trial, or all-or-none, learning perspective. This perspective was the focus of a controversy on how learning occurs, which was debated in the literature shortly before the publication of her first experiments (see Roediger & Arnold, 2012). The prevailing view was that learning is an incremental process and that with each repeated exposure associations are strengthened (e.g., Ebbinghaus 1885/1964). In

contrast, the all-or-none view assumes that an association is formed completely or not at all (e.g., Rock, 1957; Estes, 1964). Because Izawa subscribed to the all-or-none view of learning, her version of test-potentiated learning was limited to the enhanced encoding component (see Izawa, 1971). She assumed that if an item could be retrieved on an initial test, it had been learned completely so on a subsequent study trial nothing more about that item could be learned. Although she discussed how a study trial and a test trial both prevented forgetting of a learned item while the trial was on-going, she did not discuss how study and/or test trials could affect forgetting after the trials were complete. Therefore, her experiments did not speak to the possibility of an enhanced retention component of test-potentiated learning.

Izawa's (1966-1971) conceptualization of test-potentiated learning was that tests increase the effectiveness of subsequent study trials. She defined the effectiveness of a study trial as the probability that an association would be formed on that trial. This definition equates test-potentiated learning with what has been here termed the enhanced encoding component. Because of her assumptions (i.e., no learning or forgetting on test trials, no forgetting on study trials, and learning is all-or-none), she interpreted the finding that more intervening tests resulted in higher recall as indicating that tests were potentiating learning. Although this result does indicate that including more tests increases the overall effectiveness of learning in the sense that fewer study trials are required to learn the same amount of information, this result does not necessarily indicate that tests increase the effectiveness of individual study trials. That is, this result does not necessarily indicate that associations were more likely to be formed on each study trial. For instance, the same result could be found if the tests increase the retention of already-learned items. If tests increase retention during learning and thus fewer items are forgotten and need to be relearned, fewer study trials would also be needed to reach the same level of recall in

conditions with more tests. In fact, results from some studies have suggested that increased retention, rather than enhanced encoding, is the basis for the increased effectiveness of conditions with more test trials (e.g., Donaldson, 1971).

However, Izawa (1968, 1969a, 1971) did provide some evidence that tests may enhance encoding. In several of her experiments, she used conditional probability to examine the proportion of never-before recalled items that were recalled for the first time after a given study trial. Although her results were not always consistent, in general, she found that a larger proportion of never-before recalled items were recalled for the first time after study trials that had been preceded by more test trials. This result supports the new encoding component of test-potentiated learning because it suggests that more never-before recalled items were learned on study trials that were preceded by more tests. It should be noted, however, that the experiments in which Izawa was able to use conditional probability were ones in which the number of tests and lag between study trials were confounded.

Although Izawa conducted many experiments on test-potentiated learning, few researchers followed up on her findings. Of those who did, many focused on the finding that tests increase the overall efficiency of learning (e.g., Karpicke, 2009; LaPorte & Voss, 1974; Royer, 1973; Young, 1971), rather than on determining if this increased efficiency was due to enhanced encoding on study trials. For example, Royer (1973) found that subjects learned foreign vocabulary words better and faster when given the opportunity to self-test during learning. In his study, one group of subjects was given flashcards with the both the foreign language word and its English translation listed on one side and the foreign language word only listed on the other. These subjects were encouraged to use self-testing and were allowed to go through the cards in a self-paced manner until they felt they had mastered the list. This group took on average less time

to master the list than a group that could not (at least as easily) self-test during learning. This latter group was given cards that only had the foreign language word and its English translation listed on one side (with nothing written on the other side). Further, the self-testing group did better on a final test than a third group of subjects that was given the one-sided cards and was allowed to study for the same amount of time as the self-testing group had used (i.e., subjects' study times were yoked). These results suggest that self-testing increased the efficiency of the learning process, but they do not necessarily indicate that this increased efficiency was caused by test-potentiated learning.

Arnold and McDermott (2012) attempted to determine if tests increase the efficiency of learning by enhancing subsequent encoding. They used conditional probability analyses to determine if tests increase learning efficiency through the enhanced encoding component of test-potentiated learning. Subjects learned Russian-English word pairs using a sequence with one test between study periods or five tests between study periods. On each test after a restudy period, conditional probability was used to examine the proportion of never-before recalled words that were recalled for the first time. As can be seen in Figure 3, this proportion was larger in the five test condition than in the one test condition. This result suggests that the additional tests potentiated learning of the not-yet-retrieved items during subsequent study and thus provides evidence that tests increase learning efficiency by enhancing subsequent encoding.

However, interpretation of this result (Arnold & McDermott, 2012) is hindered by the same confound that was present in many of Izawa's (e.g., 1966) experiments; the number of tests and the lag between study periods was confounded. In a follow-up study, Arnold and McDermott (in preparation) attempted to unconfound these variables by including a control condition. In this experiment, subjects learned Russian-English word pairs using a sequence with five consecutive

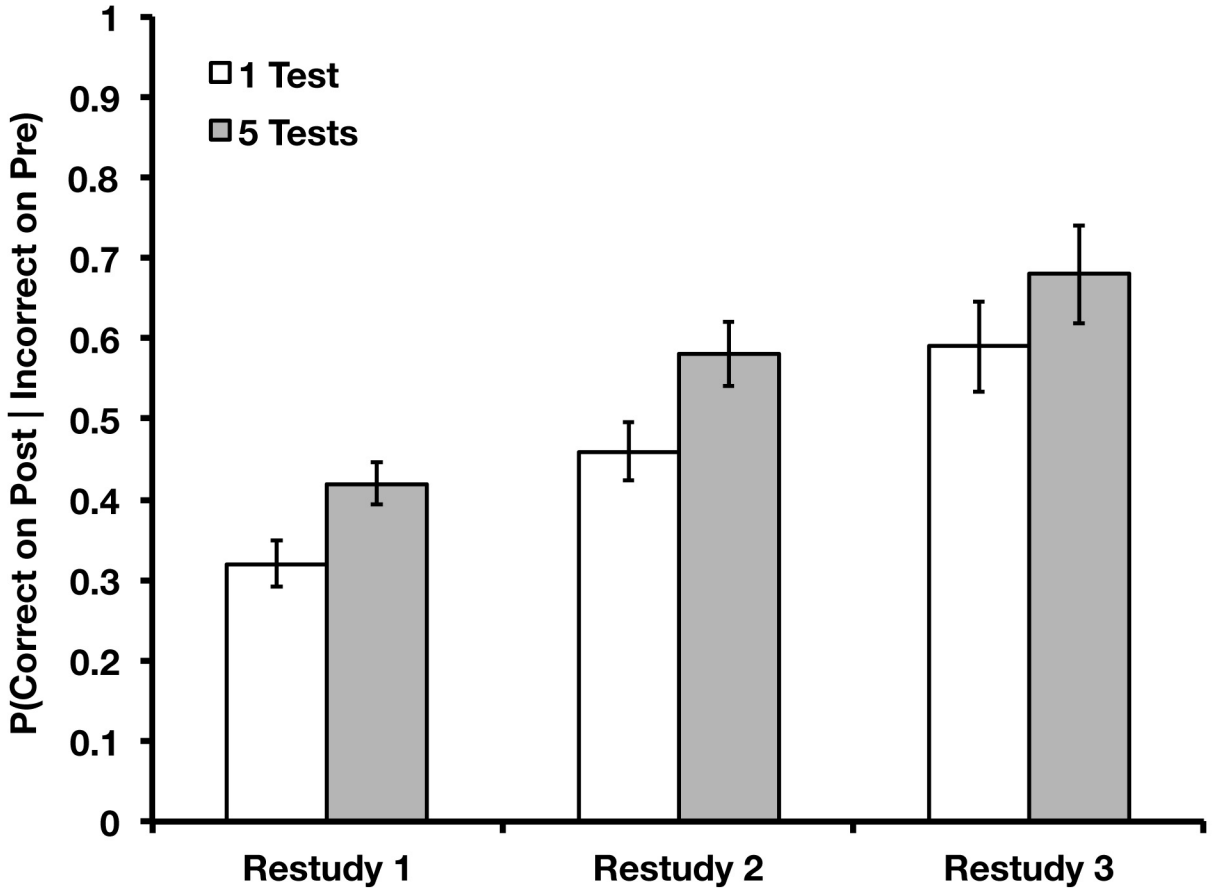

Figure 3. Mean proportion of newly retrieved items in Arnold and McDermott (2012) as a function of restudy number and the number of prior tests. Error bars represent standard errors of the mean. Adapted from their Figure 4.

tests, two consecutive tests, or two spaced tests between study periods. In the two spaced tests condition, subjects engaged in a distractor task between their two tests. This distractor task took an equivalent amount of time as three tests. In this way, the lag between study trials in the two spaced tests condition was equivalent to the lag in the five tests condition. A larger proportion of previously unrecalled words were recalled for the first time following the restudy period in the five tests condition relative to both the two consecutive tests and two spaced tests conditions. This result indicates that the enhancing effect was due to the tests and not to the lag between study periods.

Free recall test-potentiated learning

The experiments discussed thus far have all focused on the role of cued recall tests in paired associate learning. However, test-potentiated learning is not limited to cued recall tests. Free recall tests also have the capacity to enhance subsequent learning. Tulving (1967) first investigated the role of free recall tests in learning around the same time that Izawa (1966) was exploring the role of cued recall tests. He approached the question from a different perspective than Izawa and reached a different conclusion. Rather than concluding that tests increase the efficiency of learning, Tulving concluded that test and study trials have an equivalent effect on learning. He had subjects learn a list of words using different sequences of study periods and free recall tests (see Figure 4). When recall was measured as a function of the number of periods (i.e., study or tests), the sequence of study and test periods had no effect on recall. Recall was equivalent whether subjects had learned the list using a standard procedure of alternating study and test periods, a sequence of repeated study periods between tests, or a sequence of repeated tests between study periods.

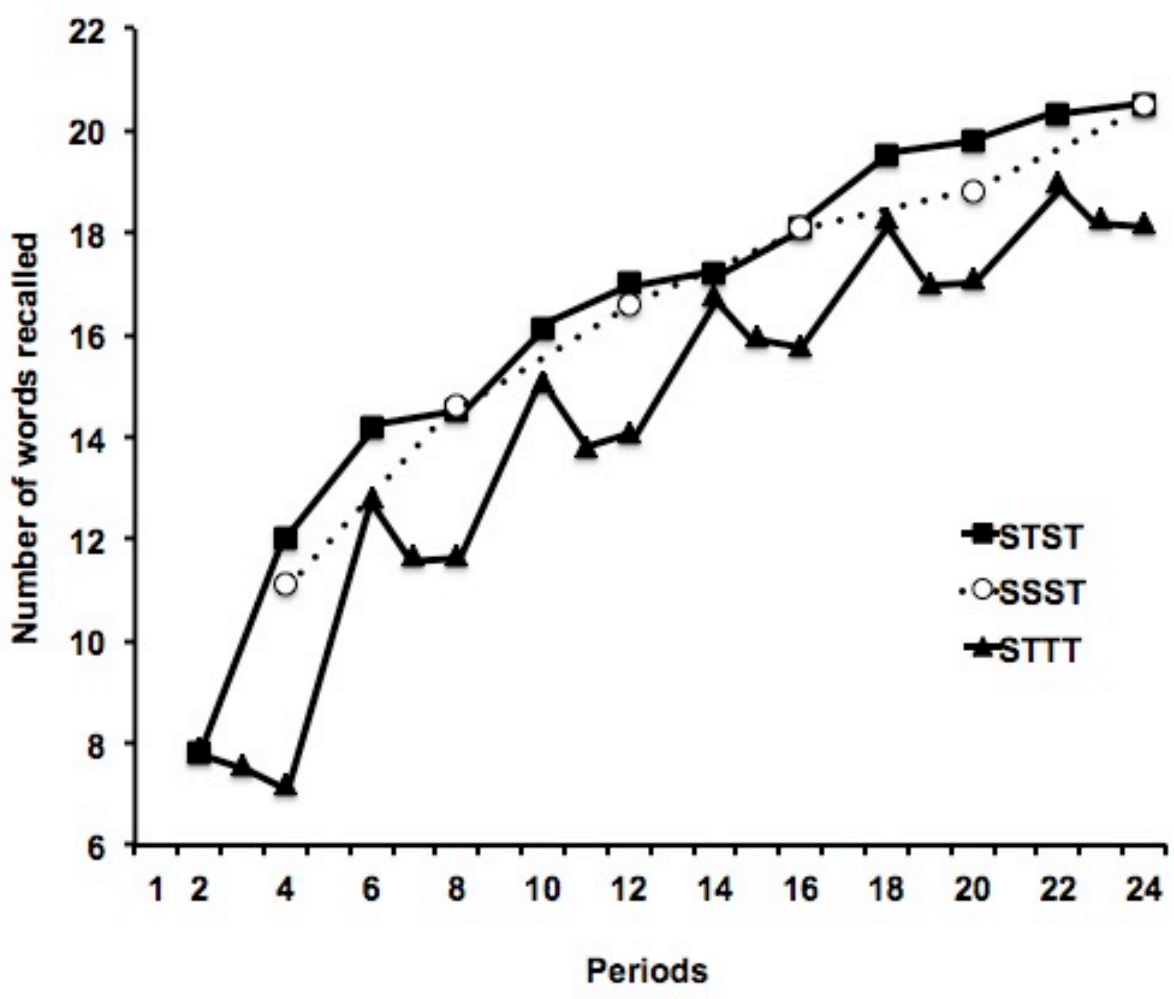

Figure 4. Mean number of words recalled on each test in Experiment 2 of Tulving (1967) as a function as a function of the number of periods and learning sequence. S = study period, T = test period. Data estimated from his Figure 2.

However, to reach this equivalence, a pattern closely resembling one that would indicate the presence of test-potentiated learning seemed to occur. In conditions with repeated tests between study periods, forgetting occurred across tests (see Figure 4). This loss was then counteracted by a greater increase in recall after study trials than in conditions without repeated tests. That is, more items seemed to be learned on study periods following repeated tests than on study periods following only one test. This suggests that the free recall tests may have potentiated learning on the subsequent study periods.

Later researchers followed up on this finding (e.g., Arnold & McDermott, 2013; Birnbaum & Eichner, 1971; Bregman & Wiener, 1970; Donaldson, 1970; Karpicke & Roediger, 2007b; Lachman & Laughery, 1968; Roediger & Smith, 2012; Rosner, 1970) and provided evidence that, contrary to Tulving's (1967) conclusion, free recall tests do increase the efficiency of learning. For example, Roediger and Smith (2012) showed that learning is hindered when no tests are included in the learning process relative to when free recall tests are included. They had subjects learn lists of words using either the standard procedure of alternating study and free recall tests or by repeatedly studying the list without intervening tests. In the repeated study condition, after a prescribed number of repeated study periods (2, 3, 4, 6, or 8) subjects were given a final test. Recall on these tests was compared to recall in the standard procedure condition as a function of the number of preceding study periods. As can be seen in Figure 5, learning was superior in the standard condition. Given the same number of study periods, more items could be recalled when tests had been taken between study periods than when tests were not included in the learning process. These results indicate that free recall tests increase the efficiency of learning. However, they do not specifically indicate whether this increased efficiency is a result of test-potentiated learning or some other effect of tests.

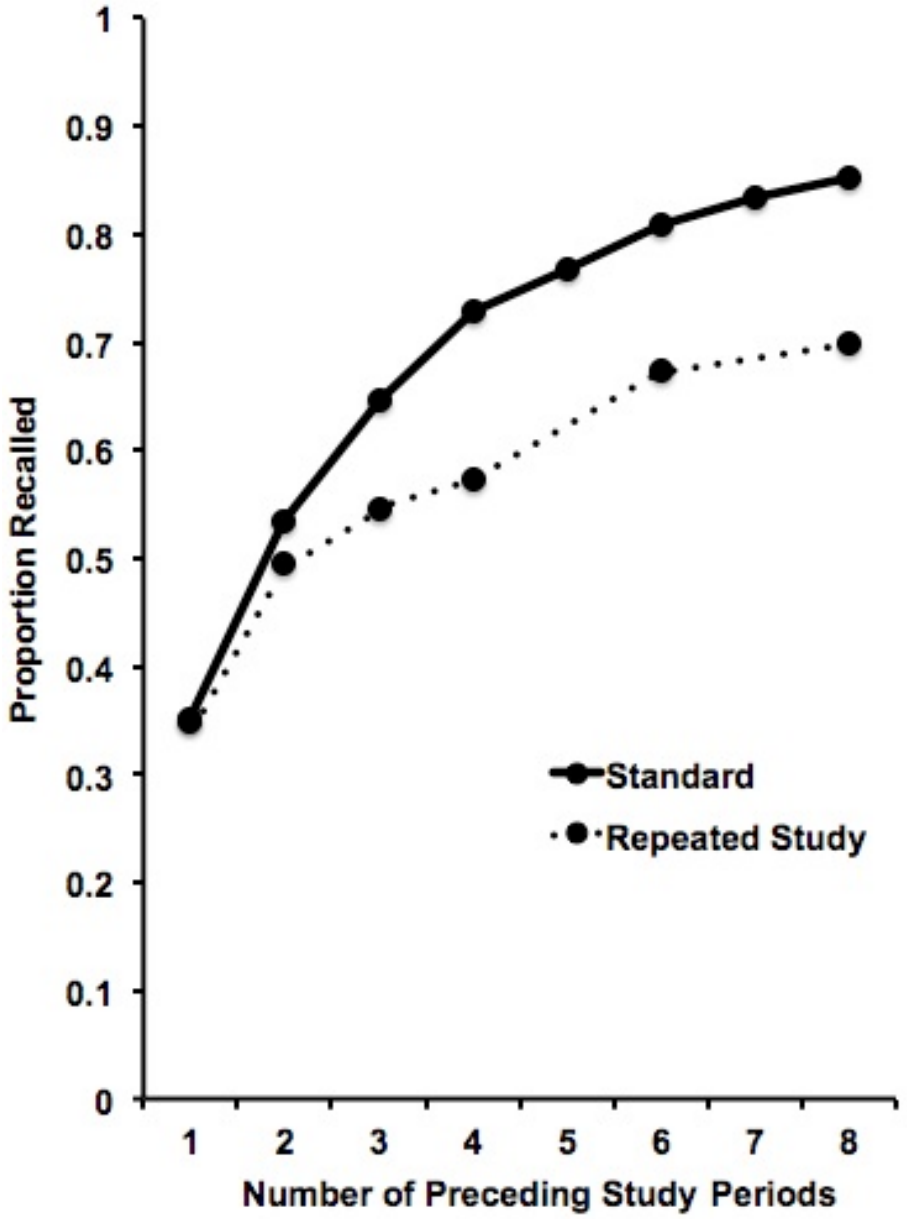

Figure 5. Mean proportion of words recalled on each test in Experiment 1 of Roediger and Smith (2012) as a function of the number of preceding study periods and learning sequence. Data obtained from second author.

Arnold and McDermott (2013) introduced a new paradigm that provides a way to measure the effects of test-potentiated learning with free recall tests and can be used to determine if increased learning efficiency is due to potentiation (see Figure 1). This paradigm was described in an earlier section. To review, half of the subjects are tested while the other half engages in a distractor task. Next, half of the subjects from each condition restudy the material. Finally, all subjects take a final test. If free recall tests potentiate learning, the difference between final recall in the restudy and no restudy conditions should be larger for subjects who had been tested than for those who had not. That is, the benefit of restudying the material should be greater after taking free recall tests then after engaging in a distractor task. As can be seen in Figure 1, this pattern was found. Subjects who took three tests prior to restudying the material gained more from the restudy opportunity than subjects who did not take these tests, thus indicating that the tests enhanced, or potentiated, learning during restudy. By including no test and no restudy control conditions, this paradigm indicates not only that tests increase the efficiency of learning, but that this increased efficiency stems from enhanced learning during subsequent study. However, this paradigm cannot distinguish between the two components of test-potentiated learning. From these results it is clear that free recall tests potentiated learning, but it is not clear whether this potentiation was due to enhanced encoding (potentiation of incorrect items) or enhanced retention (potentiation of correct items).

Summary

Together, these studies strongly indicate that both cued recall and free recall tests increase the efficiency of learning (e.g., Arnold & McDermott, 2012; Arnold & McDermott, 2013; Arnold & McDermott, in preparation; Birnbaum & Eichner, 1971; Bregman & Wiener, 1970; Donaldson, 1970; Izawa, 1971; Karpicke, 2009; Karpicke & Roediger, 2007; Lachman &

Laughery, 1968; LaPorte & Voss, 1974; Roediger & Smith, 2012; Rosner, 1970; Royer, 1973; Young, 1971). That is, by including tests between study periods, fewer study periods are needed to reach the same level of recall. The evidence for how tests increase learning efficiency, whether through test-potentiated learning or some other effect of testing, is not quite as strong. However, the evidence that does exist suggests that this increased efficiency comes at least in part from a potentiating effect (e.g., Arnold & McDermott, 2012; Arnold & McDermott, 2013; Arnold & McDermott, in preparation; Izawa, 1968, 1971; Rosner, 1970; but see Donaldson, 1970). Taking a test prior to restudying potentiates, or enhances, learning during the subsequent restudy opportunity.

For cued recall tests, this potentiating effect seems to be driven at least in part by the enhanced encoding component of test-potentiated learning (Arnold & McDermott, 2012; Arnold & McDermott, in preparation; Izawa, 1968, 1971). Prior cued recall tests enhance learning of items that could not be recalled on the initial test. Experiment 1 of this dissertation examines why cued recall tests have this effect. Whether or not the potentiating effect is also driven by the enhanced retention component of test-potentiated learning is unclear. No prior studies have specifically examined whether or not prior cued recall tests enhance the benefit of restudying for items that could be recalled on the initial test. Experiment 2 of this dissertation attempts to rectify this situation by investigating the enhanced retention component in a cued recall paradigm. For free recall tests, the effects of enhanced encoding and enhanced retention have not been separated. While the evidence suggests that free recall tests potentiate learning (Arnold & McDermott, 2013), it is not clear whether this effect is driven by potentiation of initially correct items, initially incorrect items, or both.

Generate-potentiated learning

Recently, Kornell, Hays, and Bjork (2009) introduced an effect that, at least superficially, is similar to test-potentiated learning. Using a paradigm based on a study by Slamecka and Fevreiski (1983), they found that making a failed generation attempt before studying enhanced learning. In their study, subjects learned low-associate words pairs by either first being presented with the cue only (e.g., *tide* –) and being asked to generate a related word before being presented with the entire cue-target pair (e.g., *tide – beach*) or by being presented with the whole pair first without making an initial guess. Final recall of the target items was enhanced when an initial guess had been made. That is, the generation attempt potentiated subsequent learning. This result is surprising because an initial incorrect guess could have interfered with learning the correct target (for discussions on the value of errorless learning, see Guthrie, 1952; Skinner, 1958; Evans et al., 2000), but instead making this initial guess enhanced learning.

At the surface level, generate-potentiated learning seems quite similar to test-potentiated learning. Both involve a retrieval attempt of some kind that later enhances learning. However, in test-potentiated learning an item is retrieved (or attempted to be retrieved) from episodic memory, whereas in generate-potentiated learning an item is retrieved (or attempted to be retrieved) from semantic memory (see Tulving, 1972). This distinction is important, and because of it, different processes may drive these two types of learning. In test-potentiated learning paradigms, subjects are placed in a retrieval mode by being told to specifically retrieve an item from a recent event (i.e., the initial study trial; Tulving, 1983). These instructions require subjects to attempt to retrieve an item from episodic memory. In contrast, in generate-potentiated learning paradigms, subjects are not placed in a retrieval mode. Instead, they are instructed to guess the target. This instruction essentially means that the subjects are supposed to retrieve a related item from semantic memory.

This instructional difference (i.e., whether or not retrieval mode is induced) has been shown to be a key factor in distinguishing retrieval effects from generation effects (Karpicke & Zaromb, 2010). For example, in one study by Karpicke and Zaromb, subjects first studied a set of words and then filled in word fragments that were paired with cue words (e.g., *heart – l_v_*). They filled in the word fragments by either using the first word that came to mind that fit the constraints of the fragment (generate condition) or by recalling a previously studied word that fit the constraints of the fragment (recall condition). Subjects in both conditions filled in the fragments with the target words (i.e., the originally studied words) approximately the same number of times, but on a later final test, recall of these target words was better in the recall condition than in the generate condition. These results indicate that filling in the word fragments was more beneficial for later memory when subjects were placed in a retrieval mode and retrieved items from episodic memory. This example illustrates the importance of retrieval mode, and suggests that because the retrieval mode is not present in generate-potentiated learning, it may be distinct from, albeit related to, test-potentiated learning.

Further evidence that test-potentiated learning and generate-potentiated learning are distinct types of learning comes from research suggesting that they have different boundary conditions. For example, studies by Grimaldi and Karpicke (2012) and Huelser and Metcalfe (2012) indicated that generate-potentiated learning only happens when learning semantically related items. Using the procedure developed by Kornell et al. (2009), Grimaldi and Karpicke had subjects learn either related cue-target pairs (e.g., *tide – beach*) or unrelated cue-target pairs (e.g., *pillow – leaf*). An initial generation attempt enhanced learning of the related pairs but had no effect on learning the unrelated pairs. In contrast, many of the test-potentiated learning experiments discussed earlier showed that prior tests enhance learning of unrelated pairs (e.g.,

Izawa, 1971; Arnold & McDermott, 2012). For example, in many of Izawa's (1966-1971) experiments, subjects learned pairs made up of three letter trigrams and two digit numbers (e.g., *FOJ – 84*). These stimuli were chosen specifically because of their lack of meaning and relation to each other, yet Izawa was still able to demonstrate test-potentiated learning. Similarly, Arnold and McDermott (2012) observed test-potentiated learning in a study with Russian-English word pairs (e.g., *medved – bear*), which were essentially unrelated pairs to the subjects because they were unfamiliar with Russian.

Grimaldi and Karpicke (2012) also demonstrated that delaying the presentation of the target item eliminates generate-potentiated learning (but see Hays, Kornell, & Bjork, 2012). They had subjects learn cue-target pairs either without an initial generation attempt, with an initial generation attempt followed by an immediate presentation of the pair (i.e., the standard procedure), or an initial generation attempt separated from the presentation of the pair using a block procedure (i.e., a block of generation attempts for all pairs was followed by a block of study trials for all pairs). They replicated the finding that an initial generation attempt followed by an immediate study enhanced learning relative to study alone, but they found no difference in final recall between the initial generation attempt with delayed study and study alone. Separating the generation attempt and the study trials into blocks eliminated generate-potentiated learning. In contrast, most of the previously discussed test-potentiated learning experiments used a blocked design (e.g., Arnold & McDermott, 2012; Arnold & McDermott, 2013; Izawa 1966, 1971; Tulving, 1967). That is, items were tested in one block and then restudied in a separate later block. Despite this separation, the initial tests were able to potentiate learning. These differences in boundary conditions suggest that at least some of the processes driving generate-potentiated learning and test-potentiated learning are different.

The Testing Effect

Test-potentiated learning is an indirect effect of retrieval. That is, retrieval affects later memory by modifying the subsequent effect of studying rather than directly affecting memory itself. However, as mentioned earlier, retrieval can also directly affect memory. One such direct effect, which has become known as the testing effect, is that retrieval can enhance retention by retarding forgetting (for reviews, see Roediger & Butler, 2011; Roediger & Karpicke, 2006b). This robust effect has a long history in the literature (e.g., Abbott, 1909; Gates, 1917; Spitzer, 1939; Thorndike, 1914) and has been observed with multiple kinds of tests including cued recall tests (e.g., Allen, Mahler, & Estes, 1969; Carrier & Pashler, 1992; Karpicke & Roediger, 2008) and free recall tests (e.g., Hogan & Kintsch, 1971; Roediger & Karpicke, 2006a). Additionally, it has been observed with a variety of different kinds of stimuli including word pairs (e.g., Carrier & Pashler, 1992), individual words (e.g., Hogan & Kintsch, 1971), pictures (e.g., Wheeler & Roedger, 1992), and prose material (e.g., Roediger & Karpicke, 2006a).

The testing effect is particularly robust after long retention intervals. For example, Roediger and Karpicke (2006a) found that the effect of testing as compared to repeated study increased as the retention interval increased. They had subjects study prose passages and then either restudy the passages or take a free recall test. Subjects then took a final test 5 min, 2 days, or 1 week later. As can be seen in Figure 6, after 5 min, recall was better in the repeated study condition than in the tested condition. However, after both 2 days and 1 week, this pattern reversed. Recall was now better for passages that were initially tested than for those that were restudied. This result demonstrates how the testing effect can emerge over time. Further, it shows that at long intervals testing can enhance memory even more than restudying can.

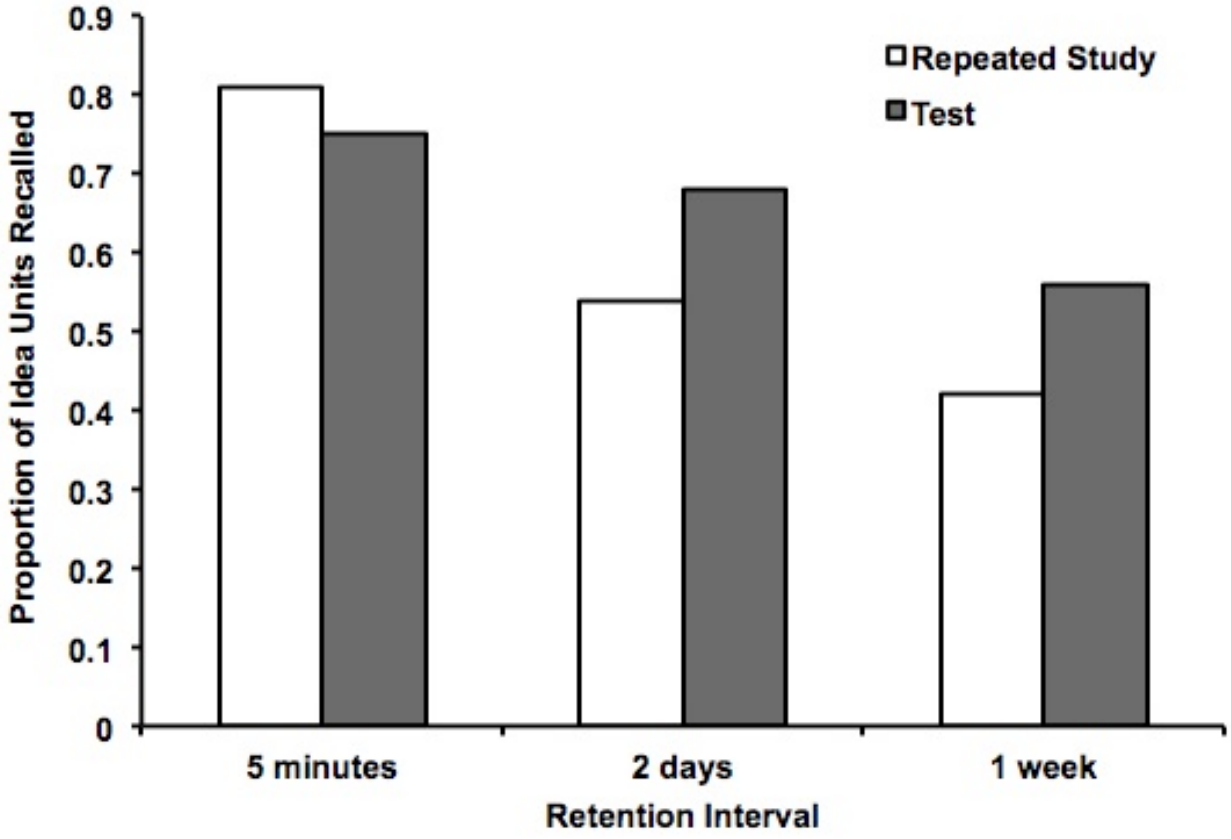

Figure 6. Mean proportion of idea units recalled in Experiment 1 of Roediger and Karpicke (2006a) as a function of retention interval and learning condition. Adapted from their Figure 1.

The testing effect is important to consider when studying test-potentiated learning. Because testing directly enhances retention, the testing effect alone could make learning more efficient. This complicates interpretation of the role of tests in learning. For instance, Roediger and Smith (2012), in an experiment described in a previous section, demonstrated that taking free recall tests between study periods enhances learning (see Figure 5). However, from their results it is unclear whether this enhancement comes from the testing effect, the test-potentiated learning effect, or a combination of both. Retrieving items on the tests may have prevented forgetting directly (the testing effect) negating the need to relearn items and thus enhancing learning. Alternatively (or in addition), tests could have potentiated learning on the study periods and thus enhanced learning indirectly (test-potentiated learning). To conclusively determine if increased learning efficiency is due to test-potentiated learning rather than to the testing effect, an experiment should be designed in such a way that learning on the subsequent study periods can be isolated. This can be done through either statistical means such as conditional probability analyses like those used in Arnold and McDermott (2012) or through control conditions such as the ones used in Arnold and McDermott (2013). In the present experiments, Experiment 1 employs conditional probability, and Experiment 2 employs conditional probabilities in conjunction with control conditions to disambiguate the effects of test-potentiated learning from the testing effect.

The Spacing Effect

Another effect with a long history in the literature is the *spacing effect*, the finding that study trials of the same item separated by some interval are better for learning than consecutive study trials of the same item (Ebbinghaus 1885/1964; Hintzman, 1974; Thorndike, 1912, Madigan, 1969; Melton, 1970; for reviews see Cepeda et al., 2006; Crowder, 1976; Dempster,

1989, 1996). Further, the farther apart study trials are spaced, the better information is learned to a point, after which increasing the spacing hurts learning. This finding is often referred to as a *lag effect*. The optimal lag between study trials depends on the retention interval, with memory after longer retention intervals benefiting from longer spaced intervals.

An experiment by Glenberg (1976) nicely illustrates the relationship between lag and retention interval. In this study, subjects studied pairs of words twice with different lags between repetitions. Between two study trials of the same item there were 0, 1, 4, 8, 20, or 40 intervening trials. In addition, pairs were tested at different retention intervals. After the last study trial of a given item, the item was tested after 2, 8, 32, or 64 intervening trials. Figure 7 shows the relationship between lag and retention interval. At long retention intervals (32 or 64 intervening trials), performance continued to increase (or reached an asymptote) as lag increased. In contrast, at short retention intervals (2 or 8 intervening trials), performance increased as lag increased at first but then declined as lag continued to increase.

Spacing test trials can also benefit learning (e.g., Landauer & Bjork, 1978; Karpicke & Roediger, 2007a). On a test trial, a subject is often re-exposed to the cue (on cued-recall tests) and the target (if retrieval is successful and/or feedback is provided), and therefore a test can be considered a type of restudy episode and can benefit from spacing. However, the effect of spacing is more complicated with tests than with study trials. In particular, the spacing of the first test following a study trial and the subsequent spacing of any remaining test trials may not have equivalent effects on learning. A test is most beneficial to learning if remembering a target item requires effortful retrieval, an observation described by the desirable difficulties framework (Bjork & Bjork, 1992). The more difficult the test, the greater the benefit of retrieval as long as

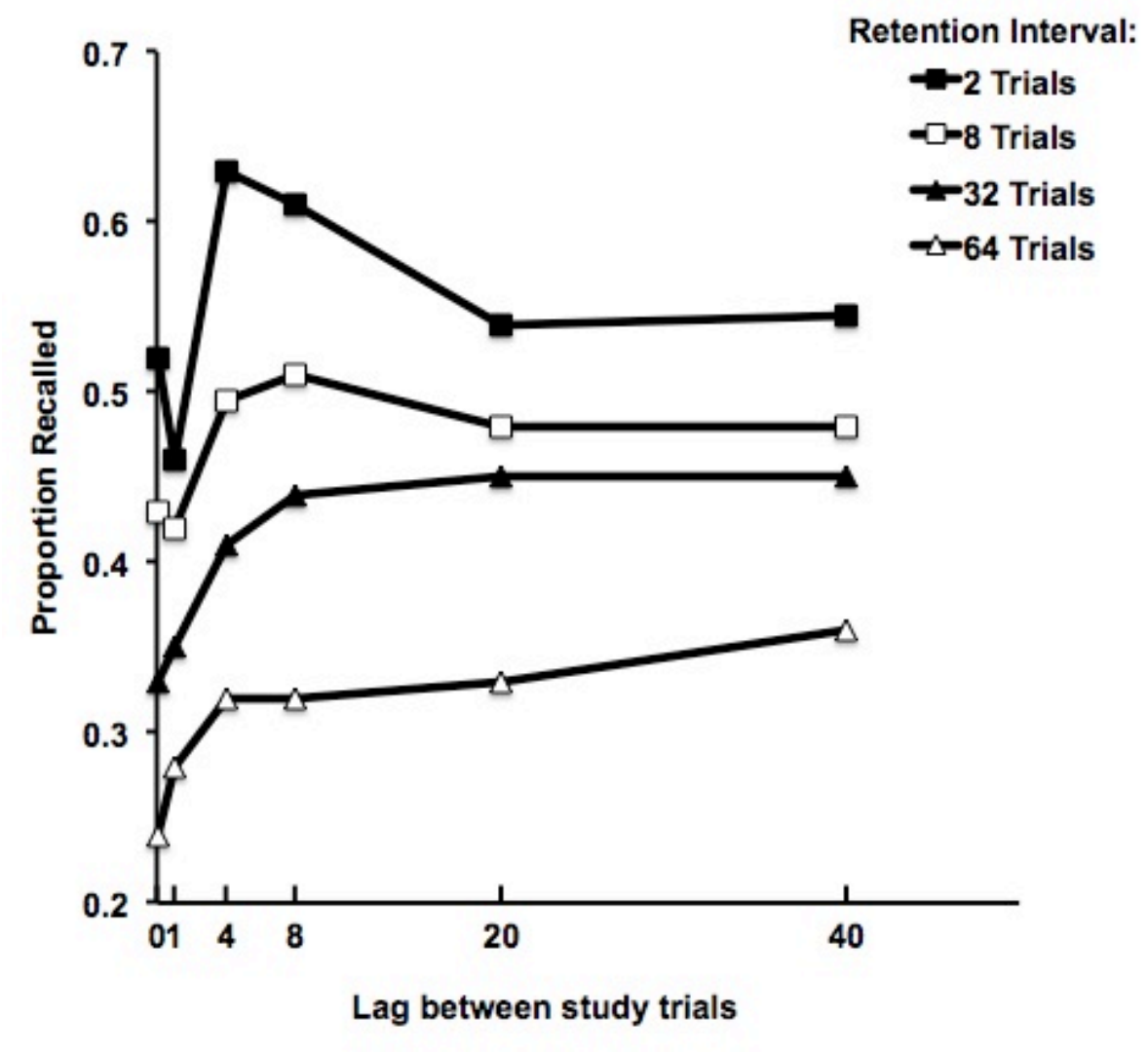

Figure 7. The proportion of items recalled in Experiment 1 of Glenberg (1976) as a function of retention interval and lag between study trials, each measured as the number of intervening trials. Data estimated from his Figure 1.

the test is not so difficult that the target item cannot be retrieved. For this reason, Landauer and Bjork (1978) argued that optimal spacing for the first test after a study trial is a relatively short lag so that the item is not forgotten before it can be retrieved and then, optimally, the lag between successive tests would gradually expand to increase the difficulty of the tests. They found evidence that this expanding retrieval sequence was better for later memory than a sequence in which all tests were spaced equally. However, more recently Karpicke and Roediger (2007a) found that expanding retrieval is better than equally spaced tests only when retention intervals are relatively short. They found that when the retention interval was longer, on the order of days, delaying the initial test enhanced learning, and the type of scheduling used to space the remaining tests had little impact on performance.

The spacing effect has important implications for understanding test-potentiated learning. In many test-potentiated learning paradigms, the number of tests between study trials is confounded by the lag between study trials (e.g., Arnold & McDermott, 2012; Izawa, 1966; Karpicke & Roediger, 2007b). This confound makes it difficult, if not impossible, to determine whether a potentiating effect is due to tests or to lag. To separate these effects, paradigms that unconfound these variables are needed (e.g., Arnold & McDermott, 2013; Arnold & McDermott, in preparation; Izawa, 1968)

However, even when paradigms are used that equate the lag between study trials, the tests themselves could be viewed as study opportunities (at least for correctly retrieved items) thereby causing the overall spacing between conditions to be unequal. For instance, Arnold and McDermott (in preparation) equated spacing between study periods while varying the number of tests. They had subjects in one condition do a distractor task (math problems) while other subjects were taking tests. Specifically, in one condition, subjects took five consecutive tests

between study periods, whereas in another condition, subjects took one test, worked on three sets of math problems, and then took a second test between study periods. Lag between study periods was equated, but in the former condition subjects were constantly engaged in the task and were re-exposed to at least some of the material three additional times while subjects in the latter condition were disengaged from the task and were not re-exposed to the material. This difference may have resulted in differential effects of spacing in these two conditions. Experiment 1 of this dissertation addresses this concern by controlling for spacing in a different way. In this experiment, in addition to equating the lag between study trials, exposure to items that are initially correctly retrieved and engagement in the task are also held constant across conditions.

Introduction to Experiments

The main objective of the present experiments was to develop a better understanding of the relation between prior retrieval and subsequent encoding. Specifically, these experiments were designed to answer two fundamental questions about test-potentiated learning: Which aspect(s) of tests enhance subsequent encoding, and do tests potentiate the effect of subsequent study on the retention of previously recalled items? Although test-potentiated learning was first introduced to the literature more than 40 years ago (Izawa, 1966), since that time research on this topic has been sparse. The few experiments that have discussed this effect have tended to do so while noting that tests increase learning efficiency, which while consistent with test-potentiated learning does not necessarily indicate that tests do potentiate learning (e.g., Karpicke, 2009; Karpicke & Roediger, 2007b; Lachman & Laughery, 1968; LaPorte & Voss, 1974; Roediger & Smith, 2013; Royer, 1973; for exceptions see Arnold & McDermott, 2012; Arnold & McDermott, 2013; Izawa, 1968, 1969a). In the present experiments, two paradigms are introduced that can potentially provide evidence that tests increase learning efficiency by potentiating learning.

There are two possible ways in which tests may potentiate learning on a subsequent restudy opportunity: by enhancing learning of initially incorrect items (enhanced encoding) and enhancing learning of initially correct items (enhanced retention). Experiment 1 examined the enhanced encoding component of test-potentiated learning, and Experiment 2 examined the enhanced retention component of test-potentiated learning. More specifically, Experiment 1 was designed to determine which aspect(s) of tests (unsuccessful retrieval, successful retrieval, or spacing) enhance subsequent encoding of initially incorrect items. This was accomplished using a paradigm that included three between-subject conditions (see Figure 8). After first studying and then taking an initial cued recall test (without feedback) on the material, a third of the subjects were given two additional cued recall tests (also without feedback) on which they were tested on all of the material again (Test All). Another third of the subjects were also given two additional cued recall tests (without feedback), but they were only tested on the items they had correctly retrieved on the initial test (Test Correct). Because recall remains relatively stable across consecutive cued recall tests (Arnold & McDermott, 2012; Izawa, 1966, 1971), these two groups of subjects were predicted to successfully retrieve the same average number of items on all three tests. For this reason, the only difference between the two groups should have been the number of items they unsuccessfully attempted to retrieve. The final third of the subjects twice restudied the items they had initially successfully retrieved instead of taking more tests (Restudy Correct). This group should have been exposed on average to the same number of complete pairs as the other two groups but did not engage in retrieval. Finally, all subjects restudied all of the material and then took a final test.

If unsuccessful retrieval enhances encoding of initially incorrect items, recall of these

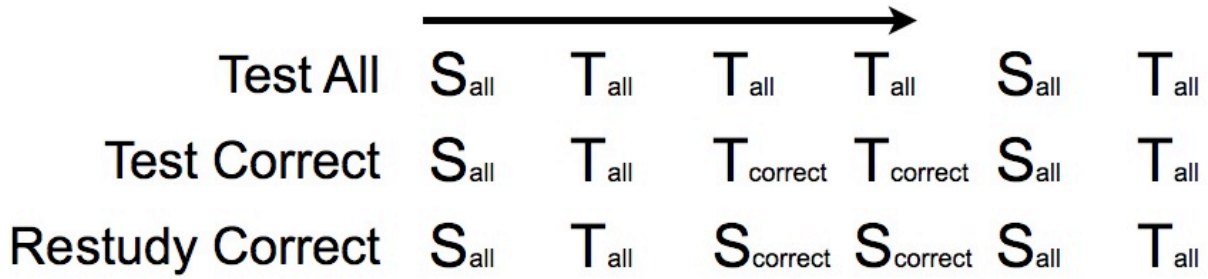

Figure 8. Design of Experiment 1. S_{all} = Study all items, $S_{correct}$ = Study initially correct items, T_{all} = Test all items, $T_{correct}$ = Test initially correct items.

items should be greater in the Test All condition than in the other two conditions. If successful retrieval enhances encoding, recall of these items should be greater in the Test All and Test Correct conditions than in the Restudy Correct condition. Finally, if spacing rather than retrieval enhances encoding, recall of initially incorrect items should be the same in all three conditions. This paradigm is a more complete test of the role of spacing in test-potentiated learning than has been previously used (e.g., Arnold & McDermott, 2012; Izawa, 1968) because subjects in all conditions were re-exposed to the same number of complete cue-target pairs and were constantly engaged in the main task. Determining whether unsuccessful retrieval, successful retrieval, or spacing enhances encoding will inform theory about what processes underlie tests enhance encoding.

Experiment 1: The Enhanced Encoding Component of Test-Potentiated Learning

The primary aim of Experiment 1 was to explore why tests enhance subsequent encoding of previously unrecalled items. Although a test is often discussed as a singular entity, it can be broken down into several components. For instance, a cued-recall test is composed of a series of trials. On some trials, subjects fail to recall the target item. On these trials, subjects are exposed to the cue only, and presumably have conducted an unsuccessful retrieval search. On other trials, participants correctly recall the target item. On these trials, participants are exposed to the complete pair and have presumably completed a successful retrieval search. Further, test trials take time and thus increase the spacing between study opportunities. If a test is taken between study periods, the study periods must be separated by at minimum the duration of the test. Any one (or combination of) these components of a test (unsuccessful retrieval, successful retrieval, spacing) may underlie subsequent enhanced encoding.

Unsuccessful retrieval

Retrieval failure may enhance subsequent learning. The process of engaging in a retrieval search for a particular target item may benefit subsequent learning of that item in several ways even when that search is ultimately unsuccessful. For instance, Kornell et al. (2009) proposed three hypotheses for how the processes engaged during a retrieval search could benefit subsequent learning: activation of related concepts could prime the target, generated items could serve as a mediator, and the correct cue-target association could be suppressed paradoxically enhancing learning. The last hypothesis stems from research showing that restudying benefits cue-target pairs more when the association between the cue and target has been suppressed prior to restudying relative to when it has not been suppressed prior to restudying (Bjork & Bjork, 1992; Storm, Bjork, & Bjork, 2008). All of these hypotheses were developed to explain generate-potentiated learning rather than test-potentiated learning. Processes underlying the enhanced encoding effect of an initial generation attempt may be different either entirely or in part from those underlying the enhanced encoding effect of an initial episodic retrieval attempt (see Karpicke & Zaromb, 2010). However, given the surface feature similarities between the two effects, at least some of the underlying processes may overlap, and therefore the processess proposed in these hypotheses may also pertain to test-potentiated learning.

Several hypotheses specific to the enhancing effect of episodic retrieval have also been proposed. One such hypothesis is that unsuccessful retrieval could enhance subsequent encoding by enhancing metacognitive knowledge (see Lachman & Laughery, 1968; LaPorte & Voss, 1974; Royer, 1973). For instance, unsuccessful retrieval may teach participants which items have not yet been learned and/or which strategies are ineffective (e.g., Gardiner & Klee, 1976; Shaughnessy & Zechmeister, 1992; Thompson, Wenger, & Bartling, 1978). The metacognitive information learned from unsuccessful retrieval can then be used during subsequent study to

make learning more efficient (e.g., Arnold & McDermott, in preparation; Metcalfe & Finn, 2008; Thiede, Anderson, & Therriault, 2003; Thomas & McDaniel, 2007).

Another way in which failed retrieval may subsequently lead to enhanced encoding is by increasing the process of remindings or study-phase retrieval (Greene, 1989; Hintzman, 1974; 2010; Thios & D'Agostino, 1976; Wahlheim & Jacoby, 2013). In the past, the role of remindings in learning has been used to explain the spacing effect. According to this theory, the effectiveness of a subsequent study opportunity depends on the degree to which it reminds the learner of the initial study opportunity. The retrieval processes engaged during the reminding enhances learning. The more effortful the retrieval process is, the greater the benefit incurred from the reminding as long as the retrieval process is ultimately successful (i.e., desirable difficulties; Bjork & Bjork, 1992). Increased spacing between study trials increases the effort needed for successful retrieval and therefore increases the benefit of the subsequent study opportunity. However, if spacing between study trials is so large that retrieval fails and the learner is not reminded of the original study opportunity, the benefit of spacing is reduced or eliminated. Tests between study periods may enhance the remindings process and, in a sense, amplify the spacing effect (Nelson, Arnold, Gilmore, & McDermott, under review).

Successful retrieval

Unsuccessful retrieval attempts are not the only component of tests that could lead to enhanced encoding. Successful retrieval could also enhance subsequent learning of initially incorrect items. That is, successfully retrieving some items could enhance the subsequent encoding of other items. There are several reasons why this may occur. For instance, successful retrieval of some items may enhance organization (Arnold & McDermott, 2013), improve

metacognitive knowledge (Postman & Schwartz, 1964), and/or reduce interference (Szpunar et al., 2008), any or all of which may enhance learning of initially incorrect items.

The importance of organization in improving learning and retention has been well documented (e.g., Bartlett, 1932; Miller, 1956; Tulving, 1962; Mandler, 1967). Organization is generally discussed and measured in studies using free recall tests. In general, as learning of a free recall list increases, organization of that list also increases (Tulving, 1962). Further, when participants are instructed to organize a set of items, they perform better on a surprise recall test than participants who are not given instructions to organize (Mandler, 1967). Zaromb and Roediger (2010) found evidence that tests enhance organization and that this may at least partially underlie the testing effect in free recall paradigms. Enhanced organization through testing may also underlie the test-potentiated learning effect (Arnold & McDermott, 2013). It may improve learning by providing a schema or structure (Barlett, 1932), which can be used to incorporate not-yet-learned items with already-learned items.

Arnold and McDermott (2013) found some evidence for this hypothesis. They had participants study a list of related items and then take three free recall tests before restudying the same items and taking a final test. They found a relationship between how organized a participant's recall was on the last free recall test before restudying and how many items they learned during the subsequent restudy period. Importantly, this relationship was not mediated by individual differences in memory ability, as measured by an independent memory task. This suggests that having better organization of already-learned items prior to restudying may enhance learning of not-yet-learned items, and therefore tests may enhance subsequent learning by improving organization of retrieved items.

Successful retrieval may also enhance learning by improving metacognitive knowledge. For instance, successfully retrieving some items (and not others) may teach participants which strategies are most effective. Many different strategies (e.g., imagery, mnemonic devices, rote memorization) can be employed when learning a list of items. Strategies vary widely in their effectiveness, and research has shown students often have a poor understanding of which strategies are most effective (Karpicke, 2009; Kornell & Son, 2009; McCabe, 2011). A test may provide participants with an opportunity to gain insight into the effectiveness of their chosen strategy or strategies. For instance, if multiple strategies are used when learning a list, participants may be able to compare the effectiveness of these strategies by noting which items were successfully recalled and what strategy had been used to learn those items. When restudying the list, this information could then theoretically be applied to the remaining items to enhance learning. This hypothesis is related to an older concept called learning-to-learn (McGeoch & Irion, 1952; Postman, Burns, & Hasher, 1970; Postman & Schwartz, 1964). This term has been used to describe nonspecific transfer across trials of the same task, or the learning of techniques and skills (such as certain strategies) from one trial that can be applied to another. Through successful retrieval on tests, participants may gain insight on how to learn and thus learn more effectively during subsequent study opportunities.

Another benefit of successful retrieval is the reduction of proactive interference (Bäuml & Kliegl, 2013; Nunes & Weinstein, 2011; Pastötter, Schicker, Niedernhuber, & Bäuml, 2011; Szpunar et al., 2008; Weinstein, McDermott, & Szpunar, 2011; Wissman, Rawson, & Pyc, 2011). Prior learning can interfere with subsequent learning, a phenomenon known as proactive interference (Postman & Keppel, 1977; Underwood, 1957). Szpunar et al. found that taking a test between studying two different lists reduced proactive inference and enhanced learning of the

subsequent list. Items within a list can also interfere with each other (Murdock, 1961; Tulving & Arbuckle, 1963). Taking a test between two study opportunities of the same list may reduce intralist interference and enhance subsequent learning of yet-to-be-recalled items. Specifically, prior successful retrieval of some items on the list could reduce interference from those items and enhance learning of the remaining items.[2]

Spacing

Finally, tests may subsequently enhance encoding by consequentially increasing the lag between study trials. Taking tests between study periods requires time and thus inherently creates a lag. As many previous experiments have shown, a lag itself can increase learning efficiency (e.g., Cepeda et al., 2006). Many experiments that have investigated test-potentiated learning have confounded the number tests between study periods with the lag between study periods (e.g., Arnold & McDermott, 2012; Izawa, 1966; Karpicke & Roediger, 2007b) and therefore have not been able to distinguish between the test-potentiated learning effect and the spacing effect. Further, studies that have not confounded these two variables have tended to use paradigms that could still create disparity in the effect of spacing. In particular, the number of tests between study periods is often confounded with re-exposure to the material and engagement with the task (e.g., Arnold & McDermott, 2013; in preparation; Izawa, 1968).

Current Experiment

Experiment 1 compares these three possible causes of the enhanced encoding component of test-potentiated learning (unsuccessful retrieval, successful retrieval, and spacing). It does so

[2] One may find a contradiction with this hypothesis and the well-established phenomenon of output interference, or the reduction in recall of items caused by earlier recall of other items from the same list (e.g., Smith, 1971). However, output interference is traditionally measured within one recall session, whereas this hypothesis is referring to the effect of recall of some items in one recall session on the probability of recalling other items during a separate recall session separated by a study opportunity.

by comparing three between-subject conditions that vary in the degree of unsuccessful and successful retrieval subjects engage in prior to restudy. Differences between conditions in the proportion of initially incorrect items retrieved after restudy will inform what kind of retrieval drives enhanced encoding. Alternatively, if there are no differences between conditions, then the results will suggest that spacing, rather than retrieval, drives enhanced encoding. Lag between study periods, re-exposure to initially correct items, and engagement in the task was equated in all three conditions, and therefore spacing should have had a similar effect across conditions.

Method

Subjects

Four hundred and sixty-seven participants completed the experiment online through Amazon Mechanical Turk in exchange for $2. Six participants were excluded from the final analyses because they indicated they were familiar with the Indonesian language. Sixty-seven participants were excluded because on a post-experimental questionnaire they indicated they had written down words during the study phases. Additionally, one participant was excluded for not following the directions. After these exclusions, 393 participants remained in the final analyses.

The remaining participants ranged in age from 18 to 71 (M = 32.5, SD = 10.9)[3]. Slightly more than half of the participants were female (n = 215). Educational background varied widely from having less than a high school degree (n = 4) to having a doctorate (n = 6). Participation was limited to those residing in the United States (as determined by Amazon), and most participants were native English speakers (n = 368). Demographic characteristics did not vary across conditions (smallest p = .25).

[3] Due to a technical error, demographic information is incomplete for fourteen participants. For three participants, information is missing for age, gender, or native English speaker status. For eleven participants, all demographic information is missing.

Design

This experiment had two parts. Part 1 was used as a baseline measure of memory ability to ensure there were no pre-experimental differences between conditions. All participants studied and then tried to recall a series of words.

Part 2 was the primary experiment and included one between-subject independent variable (treatment after initial test) with three levels (see Figure 8). After the initial test, participants took two additional tests that included all items (Test All condition; $n = 137$), took two additional tests that included only items answered correctly on the initial test (Test Correct condition; $n = 123$), or twice restudied items answered correctly on the initial test (Restudy Correct condition; $n = 133$).

Materials

For the baseline memory measure in Part 1, three lists of 15 words were taken from Roediger and McDermott (1995) for a total of 45 words. Each list contained words (e.g., *thread*, *pin*, *eye*) related to one critical item (e.g., *needle*) that never appeared in the list. Words were studied in a blocked fashion such that words from the same list were studied together. The order the lists were presented was randomized for each participant. Further, within each list, words were presented in a new random order for each participant.

In Part 2, 30 Indonesian-English word pairs (e.g., *kunci – lock*) were used as stimuli (see Appendix A). For each study and test period, items were presented in a different random order determined on a participant-by-participant basis. The word pairs were unrelated to the words used in Part 1.

On the post-experimental questionnaire, participants were given three questions answered on a Likert scale ranging from 1-5. The questions were: How interesting did you find this quiz?

[1 = not at all interesting, 5 = extremely interesting], How difficult did you find this quiz? [1 = extremely easy, 5 = extremely difficult], and How much effort did you put into trying to answer the questions? [1 = hardly any effort at all, 5 = a great deal of effort]. Additionally, participants were asked if they wrote down any words from Part 1 or Part 2 (e.g., on a piece of paper) to help on the memory test. Participants who indicated they had written words down were excluded from the analyses.

Procedure

Before beginning the experiment, participants answered a series of demographic questions. They were then told that this experiment had two parts: In Part 1, they would be learning a series of 45 words and in Part 2, they would be learning 30 Indonesian-English word pairs. Instructions were the same for all participants.

In Part 1, participants first studied all 45 words. Words were presented one at a time for 3 s with a 500 ms interstimulus interval. Next, participants were given 3 min to recall as many words as possible[4]. Participants were instructed to type each word they could remember. Recalled words appeared on the screen and remained there for the duration of the test.

After completing Part 1, participants were given additional instructions regarding Part 2. First, participants studied all 30 word pairs. Each word pair was presented individually on the screen for 5 s with a 500 ms interstimulus interval. Next, as a short distractor task to clear short-term memory (Glanzer & Cunitz, 1966) participants worked on a series of five simple addition and/or subtraction math problems involving one and two digit numbers. They had 5 s to answer each problem and there was a 500 ms interstimulus interval between trials. Participants were then given a cued recall test on all of the word pairs. Each Indonesian word was presented alone

[4] A cumulative recall curve indicated that 3 min was sufficient time for participants to recall the words they could remember.

on the screen for 5 s during which time participants were instructed to type the English translation. There was a 500 ms interstimulus interval between trials.

The procedure then varied by condition. Participants in the Test All condition completed two more tests with all 30 words pairs in the same fashion as the initial test. Participants in the Test Correct condition also completed two more tests. However, on these subsequent tests, they were only tested on items that they had answered correctly on the initial test. Participants in the Restudy Correct condition did not take additional tests. Instead, they restudied the word pairs twice. However, they only restudied word pairs that they had correctly recalled on the initial test. Each previously correct word pair was presented individually on the screen for 5 s with a 500 ms interstimulus interval.

In both the Test Correct and Restudy Correct conditions, test or restudy trials were interspersed with addition and/or subtraction problems, which replaced word pairs that had not been correctly recalled on the initial test. For each math problem, participants had 5 s to type their answer, and there was a 500 ms interstimulus interval between trials. In this way, each test or restudy period involved 30 trials: x number of test or restudy trials for each word pair that had been correctly recalled on the initial test and 30-x math problems. The test/restudy trials and math problems trials were randomly intermixed.

Next, participants in all conditions restudied all of the words pairs. As in the initial study period, each word pair was presented individually on the screen for 5 s with a 500 ms interstimulus interval. Participants were then given 5 more math problems (5 s per problem/ 500 ms interstimulus interval) before taking a final test over all 30 word pairs. The final test proceeded in the same fashion as the initial test. Finally, participants were given a post-experimental questionnaire.

Results

Results from the baseline measure given in Part 1 are presented first to determine if there were pre-experimental group differences. To preview, no group differences were found, and therefore the baseline measure was not used in subsequent analyses.

Next, results from the main experiment (Part 2) are presented. These results are divided into two sections: initial test results (including all tests given prior to the restudy period) and final test results. Finally, results from the post-experimental questionnaire are presented.

For both Experiments 1 and 2, partial eta squared (η_p^2) was used as a measure of effect size for analyses of variance (ANOVA) and Cohen's *d* was used as a measure of effect size for t-tests.

Baseline Measure

The number of words recalled during Part 1 was used as a baseline measure of memory ability. Memory ability in this context encompasses any factor unrelated to the experimental manipulation that may cause performance to vary such as but not limited to: intelligence, working memory capacity, effort, interest, and attention. This measure was included because of the unrestrained nature of conducting an online experiment with a between-subject manipulation and was used as a way to determine if the between-subject groups used in the main experiment differed pre-experimentally.

The range in the proportion of words recalled was large (min = .02, max = .98) indicating that memory ability varied widely[5]. However, a one-way ANOVA indicated that there were no significant differences between the average proportion of words recalled in the between-subject

[5] Due to a technical error, data from six participants were not included in the baseline measure analyses.

groups used in the main experiment ($M_{\text{Test All}}$ = .38, $M_{\text{Test Correct}}$ = .38, $M_{\text{Restudy Correct}}$ = .38), $F < 1$, suggesting that memory ability did not differ between groups.

Each of the three lists used in this baseline measure contained words related to one critical item (e.g., *needle*) that was not presented. This material often produces false alarms in the form of recall of these critical items (Roediger & McDermott, 1995). Recall of critical items did not differ between conditions, $F < 1$. Participants in all conditions recalled on average slightly less than one of the three (1/3) critical items ($M_{\text{Test All}}$ = .32, $M_{\text{Test Correct}}$ = .28, $M_{\text{Restudy Correct}}$ = .30). This result further confirms that there were no pre-experimental group differences.

Main Experiment

Initial test. After the initial study period, all participants took a test on all word pairs. As expected given that the experimental manipulation had not yet occurred, there were no group differences in the proportion of words recalled on this initial test (see Figure 9), $F < 1$. Participants in the Test All (M = .30), Test Correct (M = .30), and Restudy Correct (M = .29) conditions recalled an approximately equivalent number of items. Importantly, this result indicates that the baseline number of initially incorrect items was the same across conditions.

Participants in the Test All condition then took two additional tests on all word pairs. Recall varied across these three tests, $F(2, 272) = 4.64$, $p = .01$, $\eta_p^2 = .03$. More words were recalled on the second (M = .31) and third (M = .31) tests than on the initial test, $t(136) = 2.16$, $p = .03$, $d = .08$ and $t(136) = 2.60$, $p = .01$, $d = .09$, respectively. This pattern is indicative of hypermnesia, or a net gain in the number of items recalled on successive tests with no intervening study trials (Erdelyi & Becker, 1974; Payne, 1987). However, the amount of hypermnesia that occurred was rather small as indicated by the small effect sizes. There was no

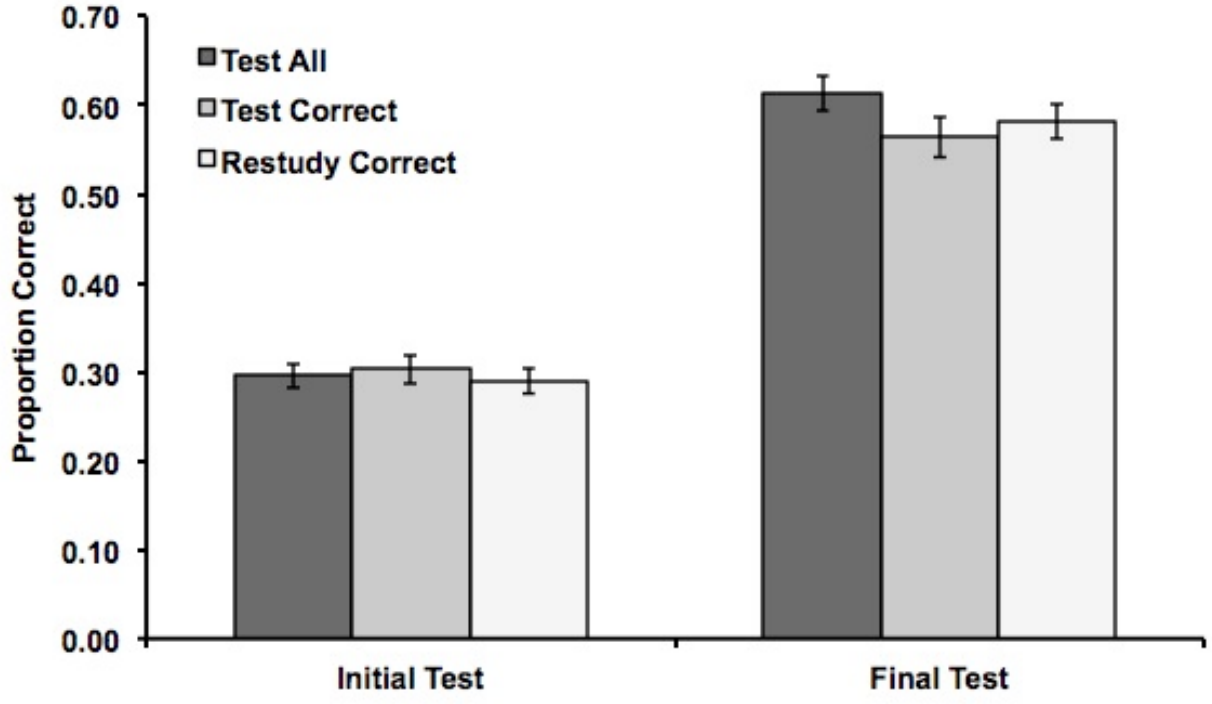

Figure 9. Mean proportion correct in Experiment 1 on the initial test and final test as a function of learning condition. Error bars represent standard errors of the mean.

difference in recall between the second and third tests, $t < 1$.

Participants in the Test Correct condition also took two additional tests. However, they were only tested on items that they had answered correctly on the initial test, and therefore they could only recall at most the same number of items on the second and third tests as on the initial test. By artificially dividing the number of items recalled on the second and third tests by 30 (the total number of word pairs studied, not the total number of word pairs tested on these subsequent tests), a one-way ANOVA can be used to compare recall from the initial test to the subsequent tests. This analysis indicates that recall dropped after the initial test, $F(2, 244) = 82.73$, $p < .001$, $\eta_p^2 = .40$. A smaller proportion of words were recalled on the second ($M = .25$) and third ($M = .26$) tests than on the initial test, $t(122) = 10.56$, $p < .001$, $d = .31$ and $t(122) = 8.98$, $p < .001$, $d = .26$, respectively. Further, more words were recalled on the third test than on the second test, $t(122) = 2.72$, $p = .01$, $d = .05$, although this difference was small as indicated by the small effect size.

An alternative way to interpret the results on these subsequent tests is as the proportion of word pairs retained from the initial test. That is, what proportion of items correctly recalled on the initial test was also correctly recalled on the second and third tests? Across both subsequent tests, retention was similar in both the Test All ($M = .78$) and Test Correct[6] ($M = .81$) conditions, $F(1, 256) = 1.38$, $p = .24$, $\eta_p^2 = .005$. Across both conditions, retention was greater on the third test ($M = .81$) than on the second test ($M = .78$), $F(1, 256) = 7.09$, $p = .008$, $\eta_p^2 = .03$. However, this increase across tests was not uniform in both conditions, as shown by a condition by test number interaction, $F(1, 256) = 4.25$, $p = .04$, $\eta_p^2 = .02$. Retention increased from the second to the third test in the Test Correct condition ($M = .79$ vs. $.83$) but did not change across tests in the

[6] Two participants in the Test Correct condition did not recall any items on the initial test and are therefore not included in this and all following retention analyses.

Test All condition (M = .78 vs. .78), $t(120) = 3.46$, $p = .001$, $d = .23$ and $t < 1$, respectively. These results indicate that overall participants in both the Test All and Test Correct conditions recalled similar numbers of initially correct items, although only participants in the Test Correct condition increased their recall of these items across tests.

Final test. Final test data can be examined in three ways: overall proportion of items recalled on the final test, proportion of initially correct items recalled on the final test, and the proportion of initially incorrect items recalled on the final test. For completeness, all three measures are discussed. However, the primary measure of interest is the proportion of initially incorrect items recalled on the final test.

First, overall recall on the final test was examined (see Figure 9). This measure includes both items that were and were not recalled on the initial test. The proportion of items recalled on the final test did not significantly differ between the Test All (M = .61), Test Correct (M = .56), and Restudy Correct (M = .58) conditions, $F(2, 390) = 1.52$, $p = .22$, $\eta_p^2 = .008$. Given that the final test occurred within the same session as the rest of the experiment, a lack of an overall advantage of the two testing conditions over the restudy condition is not surprising. Previous research has shown that given short retention intervals, restudying can lead to as good if not better final test performance than retrieval practice (e.g., Roediger & Karpicke, 2006a).

Another way to analyze final test data is by comparing it to the proportion of words that were recalled on the initial test using a 2 (initial test, final test) X 3 (condition) mixed ANOVA. As in the one-way ANOVAs for both the initial and final test data, this analysis revealed no effect of condition, $F < 1$. Collapsing across conditions, the proportion of items participants could recall increased from the initial test to the final test (M = .30 vs. .59), $F(1, 390) = 1365.91$, $p < .001$, $\eta_p^2 = .78$. However, this main effect was qualified by a significant test by condition

interaction, $F(2, 390) = 4.26$, $p = .02$, $\eta_p^2 = .02$, suggesting that this increase was not uniform across conditions. Visual inspection of Figure 9 suggests that the interaction emerged because, at least numerically, recall on the initial test in the Test All condition was slightly lower than recall in the Test Correct condition, whereas this pattern reversed on the final test. However, given that the one-way ANOVAs suggest that there were no significant differences across conditions on either the initial test or final test, this difference may not be meaningful.

Next, the proportion of initially incorrect items that were recalled on the final test was examined. This is the primary measure of interest and will be referred to as the proportion of newly retrieved items (i.e., items that were retrieved for the first time on the final test). If enhanced encoding is due to repeated unsuccessful retrieval, the proportion of newly retrieved items should be greater in the Test All condition than in both the Test Correct and Restudy Correct conditions. In contrast, if the enhanced encoding component of test-potentiated learning is caused by repeated successful retrieval, the proportion of newly retrieved items in the Test Correct and Test All conditions should be equivalent and both should be greater than the proportion recalled in the Restudy Correct condition. Finally, if enhanced encoding is due to spacing, the proportion of newly retrieved items should be equal in all three conditions.

As can be seen in Figure 10, the proportion of newly retrieved items varied by condition, $F(2, 390) = 4.58$, $p = .01$, $\eta_p^2 = .02$. More items were newly retrieved in the Test All ($M = .54$) condition than in both the Test Correct ($M = .46$) and Restudy Correct ($M = .48$) conditions, $t(258) = 2.76$, $p = .01$, $d = .34$ and $t(268) = 2.35$, $p = .02$, $d = .29$, respectively. There was no difference between the Test Correct and Restudy Correct conditions, $t < 1$. These results support the hypothesis that the enhanced encoding component of test-potentiated learning is caused by

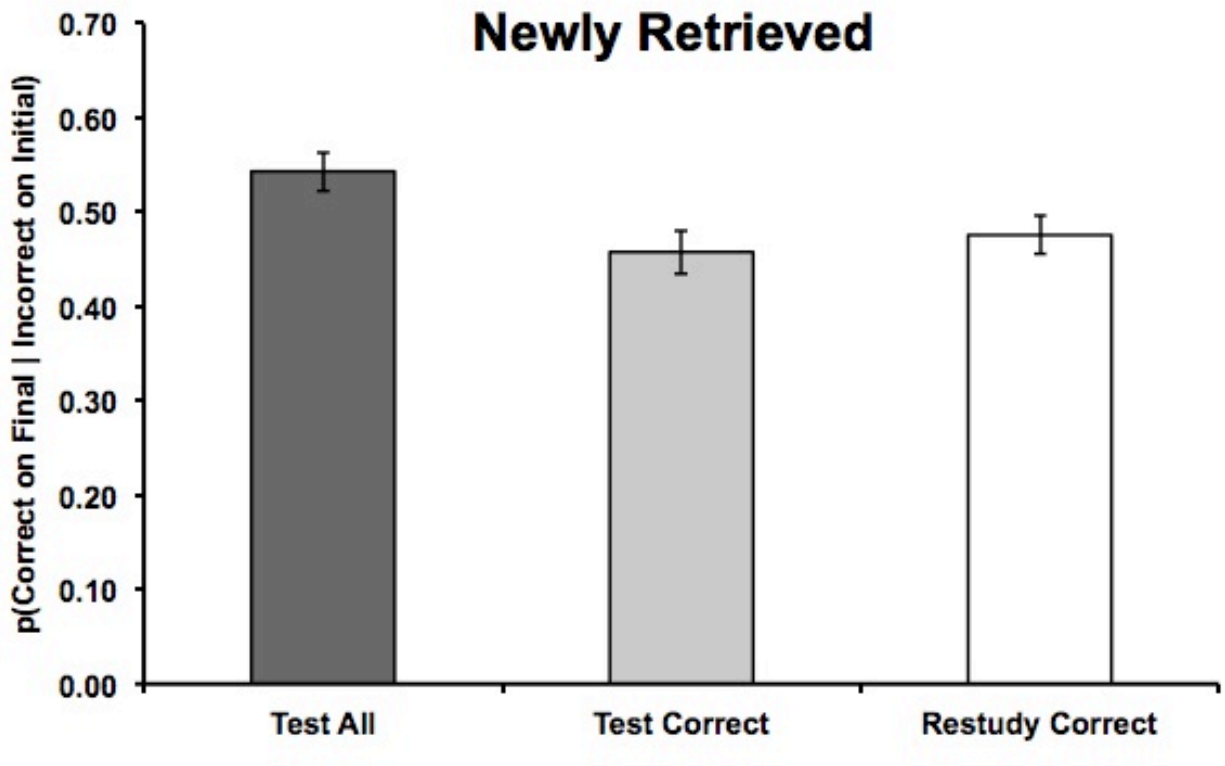

Figure 10. Mean proportion of items newly retrieved on the final test in Experiment 1 as a function of learning condition. Error bars represent standard errors of the mean.

previous unsuccessful retrieval attempts of the to-be-encoded items.

One could argue that the pattern described above may not have been due to enhanced encoding during the restudy phase in the Test All condition relative to the other two conditions, but may have instead been due to the two additional opportunities participants in the Test All condition had to correctly retrieve the initially incorrect items. That is, the pattern of results that appears to be due to enhanced encoding caused by repeated unsuccessful retrieval may actually be due to successful retrieval of the initially incorrect items on one or both of the subsequent tests. Supporting this viewpoint is the finding that in the Test All condition the proportion of items recalled increased after the initial test.

To examine this hypothesis, a more stringent analysis was conducted in which a newly retrieved item was defined as one that was retrieved on the final test and was not retrieved on any previous test (rather than defined as one that was retrieved on the final test and was not retrieved on the initial test). This definition reduces the proportion of newly retrieved items in the Test All condition but does not affect the proportion of newly retrieved items in either the Test Correct or the Restudy Correct conditions. Although this definition provides a more stringent test of the hypothesis, it also complicates comparisons between conditions for two reasons. First, it reduces the denominator in the Test All condition but does not affect the denominator in the Test Correct and Restudy Correct conditions. Using this definition, fewer items were available to be learned in the Test All condition than in the other conditions making comparisons across conditions of conditional probability difficult to interpret. Second, the initially incorrect items that subjects in the Test All condition recalled on the second and third tests were by definition easier items for those subjects than the initially incorrect items they did not recall on those subsequent tests. This creates what has been termed an item selection artifact, or, more appropriately, an item-by-

subject selection artifact (for a discussion on item-by-subject selection artifacts, see Roediger & Arnold, 2012). Calculating newly retrieved items in this new way removes the easiest initially incorrect items for subjects in the Test All condition leaving only the more difficult items available to be newly retrieved, whereas all initially incorrect items, both easy and difficult, are available to be newly retrieved in the other two conditions. This second complication is particularly detrimental and makes interpretation of this analysis particularly difficult if not impossible.

Given this more stringent definition of newly retrieved items, there are no longer significant differences across conditions, $F(2, 389) = 1.51$, $p = .22$, $\eta_p^2 = .008$. Participants in the Test All ($M = .51$) condition and the Test Correct ($M = .46$) and Study Correct ($M = .48$) conditions all recalled approximately the same proportion of previously incorrect items. Further, when recall of initially incorrect items is collapsed across the Test Correct and Study Correct conditions ($M = .47$) and compared to recall in the Test All condition, the differences between conditions are still not significant, $t(390) = 1.64$, $p = .10$, $d = .17$. These results suggest that the additional opportunities in the Test All condition to subsequently retrieve initially incorrect items may have contributed to the larger proportion of newly retrieved items observed in this condition. However, given the inherent complications in this analysis (conditional probabilities with different denominators and the item-by-subject selection artifact), these results do not negate the possibility that repeated unsuccessful retrieval led to enhanced encoding in the Test All condition.

Finally, the proportion of initially correct items that were retained (i.e., recalled on the final test) was examined. There were significant differences in retention across conditions (see Figure 11), $F(2, 388) = 4.51$, $p = .01$, $\eta_p^2 = .02$. More items were retained in the Restudy Correct condition ($M = .91$) than in the Test All condition ($M = .85$), $t(268) = 3.07$, $p = .002$, $d = .37$.

This difference was likely due to the additional study opportunities in the Restudy Correct condition. As mentioned earlier, previous research has shown that given short retention intervals like the one in the present experiment, restudying can lead to greater final recall than testing (e.g., Roediger & Karpicke, 2006a). No other differences were significant ($M_{\text{Test Correct}} = .89$).

Post-experimental questionnaire. On the post-experimental questionnaire, participants were asked to rate the degree to which they were interested in the experiment, how difficult they found the experiment, and the amount of effort they put into the experiment (see Table 1). A one-way ANOVA was used to look for differences across conditions for responses to each of these questions. No differences were found in how interesting participants rated the experiment or in how much effort they reported exerting, $F(2, 389) = 2.38$, $p = .09$, $\eta_p^2 = .01$ and $F(2, 386) = 1.06$, $p = .35$, $\eta_p^2 = .005$, respectively. There was a marginal difference in difficulty ratings across conditions, $F(2, 388) = 2.88$, $p = .06$, $\eta_p^2 = .01$, with a trend suggesting that the more items that were tested prior to restudying, the more difficult participants found the experiment ($M_{Test\ All} = 4.28$, $M_{Test\ Correct} = 4.15$, $M_{Restudy\ Correct} = 4.05$).

Discussion

This experiment was designed to determine why previous tests enhance subsequent encoding of tested (but previously unretrieved) items. Three hypotheses were tested: Enhanced encoding may be due to repeated unsuccessful retrieval of the target items, repeated successful retrieval of other items, or the spacing between study periods that intervening tests create. The results support the hypothesis that repeated unsuccessful retrieval attempts for the target items causes enhanced encoding of those items during subsequent restudy. The proportion of newly retrieved items was greater when participants were given multiple tests on all items (Test All condition) than when given multiple test or restudy trials on only items they had previously

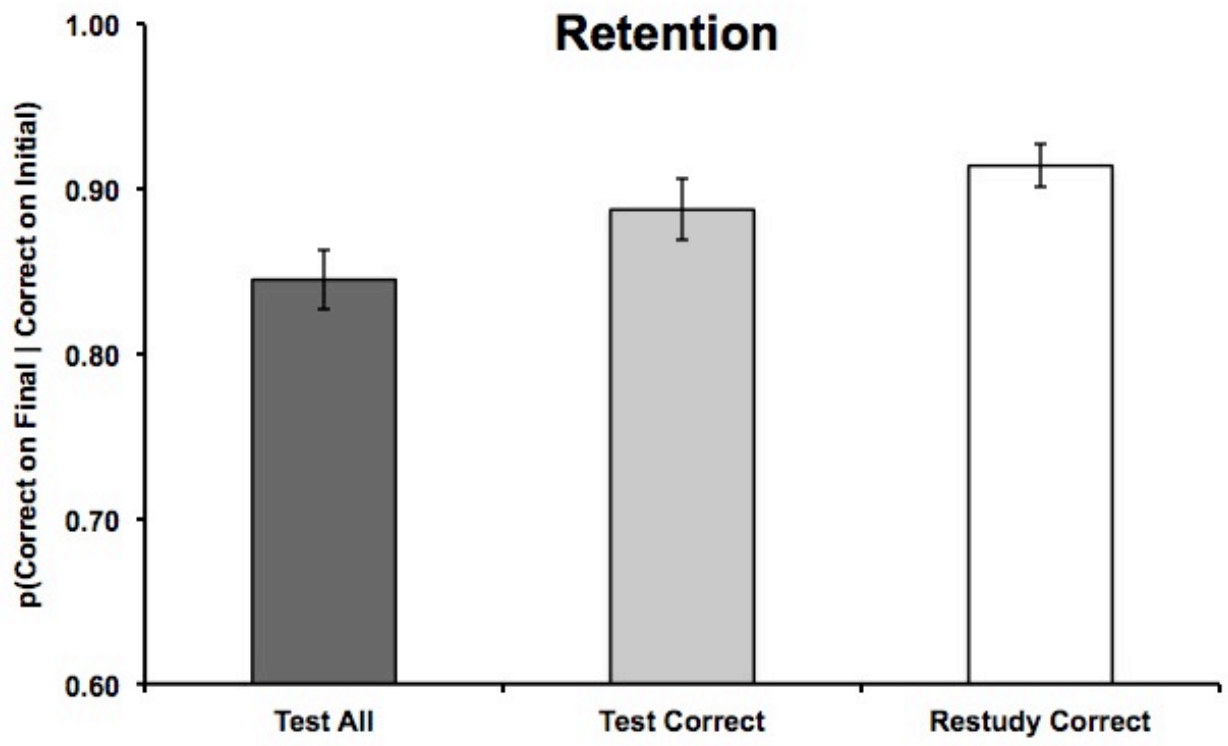

Figure 11. Mean proportion of items retained from the initial test to the final test in Experiment 1 as a function of learning condition. Error bars represent standard errors of the mean.

Table 1

Mean responses from the post-experimental questionnaires in Experiments 1 and 2. Responses were made on a Likert scale ranging from 1 (not interesting, difficult, or effortful) to 5 (very interesting, difficult or effortful).

			Session 1			Session 2		
Experiment	Learning Condition		Interest	Difficulty	Effort	Interest	Difficulty	Effort
Experiment 1		Test All	3.81	4.28	4.68	-	-	-
		Test Correct	4.07	4.15	4.76	-	-	-
		Restudy Correct	4.00	4.05	4.76	-	-	-
Experiment 2	3 Tests	Restudy	3.71	3.95	4.66	3.84	4.13	4.27
		No Restudy	3.57	4.09	4.57	3.81	4.14	4.33
	1 Test	Restudy	3.81	3.73	4.60	3.72	4.30	4.35
		No Restudy	4.05	3.95	4.76	3.82	4.37	4.38

successfully retrieved (Test Correct and Restudy Correct conditions). Although this finding did not survive analyses with the more stringent definition of newly retrieved items, given the inherent problems in this alternative definition (i.e., differential denominators, item-by-subject selection artifact), this null result does not invalidate the original finding.

These results inform what mechanisms may underlie the enhanced encoding component of test-potentiated learning. According to these results, processes that occur during unsuccessful retrieval attempts drive subsequent enhanced encoding[7]. For example, possible processes include activation of partial knowledge during the retrieval attempt (Kornell et al., 2009), improved metacognitive knowledge gained from unsuccessful retrieval attempts (Arnold & McDermott, in preparation), or enhanced retrieval processes engaged during the restudy phase as a result of previous unsuccessful retrieval (Nelson et al., under review).

Kornell et al. (2009) proposed three hypotheses for why a failed generation attempt may potentiate subsequent learning. Although different processes may underlie generate-potentiated learning and test-potentiated learning, there are enough similarities between the two effects to consider how these hypotheses may (or may not) pertain to test-potentiated learning. The first hypothesis they proposed was that the retrieval search could activate related concepts, which in turn could enhance learning of the target item via priming through spreading activation (Collins & Loftus, 1975). For example, if a participant is given the cue word *tide* and asked to guess the semantically related target word (*beach*), related concepts such as waves, ocean, and surf may be

[7] Due to the lack of differences between conditions when New Learning is defined stringently, this conclusion should be interpreted cautiously. Conclusions here are made using the original definition of New Learning because of the problems inherent in interpreting the stringent definition (i.e., differential denominators, item-by-subject selection artifact). However, the present experimental design cannot resolve the issue that hypermnesia may have at least partially contributed to the advantage in the Test All condition.

activated even if *beach* is not retrieved. Activation of these related concepts could prime *beach* and therefore enhance encoding during a subsequent study opportunity.

Grimaldi and Karpicke's (2012) finding that engaging in an initial generation attempt enhanced learning for semantically related pairs but not for semantically unrelated pairs supports the priming hypothesis (see also Huelser & Metcalf, 2012). For an unrelated pair, a generation attempt would be unlikely to activate concepts related to the target item and therefore priming of the target would be unlikely to occur. Further, even when pairs were related, delaying the study opportunity eliminated the advantage of the initial generation attempt. Because priming tends to dissipate quickly (i.e., after only a few seconds or intervening items; McNamara, 1992; Neely, O'Connor, & Calabrese, 2010), delaying the study opportunity would have eliminated any effect of priming. Because the initial generation attempt enhanced subsequent learning only under conditions in which priming would have been present (related items, immediate study), these results indicate that priming may drive the effect.

However, in other experiments (including Experiment 1) enhanced encoding has been observed for unrelated pairs when an initial episodic retrieval attempt (rather than an initial generation attempt) is made (Arnold & McDermott, 2012; Izawa, 1971). Enhanced encoding from testing has been observed with unrelated pairs such as two digit-three syllable pairs (Izawa, 1971) and foreign language word-English translation pairs (which are seemingly unrelated to participants who have no previous knowledge of the foreign language; Arnold & McDermott, 2012). These results suggest that priming may not underlie the enhanced encoding effect of episodic tests.

Further, enhanced encoding has been observed when the lag between the retrieval attempt and subsequent study trial has been delayed (Arnold & McDermott, 2012; Hays et al., 2012;

Izawa, 1971). This delayed enhancing effect has been found following both an initial episodic retrieval attempt (Arnold & McDermott, 2012; Izawa, 1971) and an initial generation attempt (Hays et al., 2012). Hays et al. found that when participants on the final test were asked to recall the cue word, rather than the target word, recall was better for word pairs in which there had been an initial generation attempt relative to trials without an initial generation attempt, even when the lag between generation and the subsequent study opportunity was delayed by several minutes and multiple intervening items. This result suggests that priming alone may not be able to account for the enhanced encoding effect following either an episodic retrieval or generation attempt.

A related explanation for enhanced encoding suggested by Kornell et al. (2009) is that an incorrect guess could serve as a mediator or additional retrieval cue for the target item. Previous research has shown that mediators that connect a cue with the target can enhance learning (Pyc & Rawson, 2010) and that a previous retrieval attempt can be an effective mediator (Soraci et al., 1999). An unsuccessful retrieval attempt may result in the retrieval of a related item that could be used as a mediator and thus enhance subsequent learning of the correct target. Because the unsuccessful retrieval attempt is more likely to produce an effective mediator if the generated item is semantically related to the target item, the finding that an initial generation attempt only enhances encoding for related word pairs (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012) is consistent with this explanation. However, because enhanced encoding following an episodic retrieval attempt has been observed for unrelated pairs (Arnold & McDermott, 2012; Izawa, 1971), mediators alone are unlikely to explain the enhanced encoding effect that follows episodic retrieval. Retrieval attempts for unrelated pairs are unlikely to produce an effective mediator (see Pyc & Rawson, 2010, 2012 for a discussion on mediator effectiveness).

The third hypothesis proposed by Kornell et al. (2009) is that unsuccessful retrieval searches suppress the cue-target association, which paradoxically enhances later learning of the cue-target pair (see also Hays et al., 2012). Unsuccessful retrieval searches activate incorrect semantic associates, thereby strengthening the association between the cue and these incorrect associates and simultaneously suppressing the association between the cue and the correct target (Anderson, Bjork, & Bjork, 1994, 2000). Previous research has demonstrated that suppressing an association prior to a restudy episode can enhance learning (Bjork & Bjork, 1992; Storm, Bjork, & Bjork, 2008), suggesting that suppression occurring during failed retrieval searches may underlie the enhanced encoding component of test-potentiated learning. Temporary suppression can last for at least 20 min (Anderson et al., 1994) so, unlike priming, suppression could explain why enhanced encoding can be observed given delays of several minutes between testing and study trials (Hays et al., 2012). However, this hypothesis may not be able to explain why unsuccessful episodic retrieval attempts can enhance subsequent learning of unrelated pairs (such as the Indonesian-English word pairs used in Experiment 1). An association between the cue and unrelated target item would be unlikely to be suppressed during a failed retrieval attempt because there is not likely to be a pre-existing association between these two items. The only way in which an association could be suppressed during failed retrieval is if during initial study some association between the cue and target was formed. That is, for association suppression to explain enhanced encoding in unrelated pairs, an assumption would have to be made that on trials in which not enough of an association between the cue and target was formed during the initial study to result in successful retrieval, enough of an association was still formed so that unsuccessful retrieval could result in suppression of that association.

Another way in which unsuccessful retrieval could potentiate learning is by improving metacognitive knowledge. Arnold and McDermott (in preparation) found support for this hypothesis by demonstrating that taking prior tests affected subsequent study decisions, a behavior based on metacognitive judgments (Metcalfe & Finn, 2008). Specifically, Arnold and McDermott found that the number of tests taken prior to a restudy opportunity (pre-restudy tests) affected how long participants chose to restudy not-yet-learned items. Participants first studied a set of Russian-English word pairs and then took either two or five sequential cued-recall tests (without feedback). They then restudied each item in a self-paced manner before taking a final test. There were no differences in how long participants chose to study items that they had correctly recalled on the pre-restudy tests. However, participants who took five pre-restudy tests chose to study initially incorrect items for significantly longer (M = 7.6 s) than participants who took only two pre-restudy tests (M = 4.9 s). This result suggests that the additional pre-restudy tests altered participants' metacognitive judgments, specifically their metacognitive judgments related to not-yet-retrieved items. These changes in metacognitive judgments affected how participants approached the restudy opportunity indicating that metacognition may at least in part underlie the enhanced encoding component of test-potentiated learning.

One way in which tests could affect metacognition is by helping participants distinguish between items they know and items they do not know (Gardiner & Klee, 1976; Thompson et al., 1978). Indeed, research has shown that testing allows participants to better predict future performance (i.e., which items they will be able to recall) relative to restudying (e.g., Glenberg, Sanocki, Epstein, & Morris, 1987; King, Zechmeister, Shaughnessy, 1980; Shaughnessy & Zechmeister, 1992). Students appear to be aware of this benefit of testing; when they use self-testing during study periods, they report doing so as a way to identify which items need

additional study (Kornell & Bjork, 2007; Kornell & Son, 2009). This information can be used to enhance subsequent encoding and improve learning (Metcalfe & Finn, 2008; Thiede et al., 2003; Thomas & McDaniel, 2007).

However, using testing as a means to identifying unlearned items does not seem to be driving the study duration differences observed by Arnold and McDermott (in preparation). In a follow-up study, after taking two or five tests, participants were asked whether they had recalled the target item for each pair on the last test they took. Participants in both conditions were highly accurate ($M_{2\ Tests}$ = 94%; $M_{5\ Tests}$ = 96%), and there were no differences between conditions. Other research has shown that after just one test, participants are generally able to accurately distinguish between items they did and did not recall (Battig, Allen, & Jensen, 1965; Gardiner & Klee, 1976). These results suggest that the additional study time used by participants who had taken more prior tests was not driven by an increased ability to distinguish between previously recalled and unrecalled items. Rather, this difference was likely driven by some other change in metacognitive knowledge. For instance, the additional tests may have given participants the additional experience needed to better distinguish between effective and ineffective strategies (e.g., Pyc & Rawson, 2012). Alternatively, the additional study time may have been driven by a feeling of frustration and thus increased determination and effort to learn the word pairs. However, it should be noted that the additional study time might not have driven the observed potentiating effect. Previous research has shown that increased self-paced study time does not always result in an increase in performance (i.e., a labor-in-vain effect; Nelson & Leonesio, 1988). The observed increase in study time could have been a consequence rather than a cause of test-potentiated learning.

Finally, unsuccessful retrieval may enhance subsequent encoding by increasing remindings. Taking tests prior to restudying may increase the engagement of retrieval processes during subsequent restudy. Nelson et al. (under review) found evidence for this hypothesis using functional magnetic resonance imaging (fMRI), which provided a way to measure responses during study trials. They scanned participants as they studied 126 low-associate word pairs. Then, during phase 2, participants were tested on 1/3 of the pairs (without feedback) and restudied another 1/3 of the pairs. The remaining 1/3 of the pairs were control items and were not tested or restudied. In phase 3, participants were scanned again as they restudied all of the word pairs. Nelson et al. compared the hemodynamic activity during the restudy trials to activity during the initial study trials as a function of what happened during phase 2. From this analysis, they found evidence indicating that during restudy in phase 3 retrieval processes were engaged and that these processes were engaged to a greater degree for previously tested items than for previously restudied or control items. Specifically, they found that activity in the left posterior inferior parietal lobule (pIPL) was greater for tested items during restudy than during initial study and that, during restudy, activity in this region was greater for tested items than restudied or control items. Activity in this region has previously been associated with successful recognition (McDermott, Szpunar, Christ, 2009; Nelson et al., 2010), and therefore these results suggest that, even though there were no explicit instructions to retrieve, retrieval processes were engaged during restudy and they were engaged to greater degree for previously tested items than for previously restudied or control items. Further, Nelson et al. (under review) found that activity in the pIPL was correlated with the amount of learning that occurred during restudy. Greater activation in the pIPL during restudy was correlated with a larger proportion of previously

unrecalled items that were recalled on a final test. Although this result is only correlational, it suggests that engaging in retrieval processes during restudy enhanced learning.

The results from the present experiment have narrowed the possible theories that can explain the enhanced encoding component of test-potentiated learning to ones that suggest that unsuccessful retrieval is the driving force behind potentiation, at least in paired-associate learning (but see Footnote 7). Theories that explain test-potentiated learning as being due to prior successful retrieval (e.g., enhanced organization, learning-to-learn, reduction in proactive interference) or theories that explain test-potentiated learning as being due to spacing cannot explain these results, thus leaving only theories that explain test-potentiated learning as being due to prior unsuccessful retrieval as viable explanations. Although this experiment was not designed to discriminate between theories that depend on prior unsuccessful retrieval, these results in conjunction with prior work do suggest some theories are more likely than others to be able to account for the results related to the enhanced encoding component of test-potentiated learning. The three theories proposed by Kornell et al. (2009) to explain generate-potentiated learning (target primed through spreading activation, incorrect guess serves as a mediator, and unsuccessful retrieval suppresses the cue-target association) are not likely candidates for viable theories of test-potentiated learning because the effect was found when using unrelated materials with a delay between the prior retrieval attempt and the subsequent encoding opportunity. Two other theories, enhanced metacognitive knowledge and increased remindings, are more consistent with Experiment 1 results. Further research is needed to determine which (if either) of these theories describe processes that underlie the enhanced encoding component of test-potentiated learning.

Prior tests may also enhance subsequent learning of initially correct items. If this enhanced retention component of test-potentiated learning exists, it may or may not be driven by the same processes as the enhanced encoding component. Experiment 2 looked for evidence of enhanced retention to determine if it is a component of test-potentiated learning. No previous studies have directly examined this possibility, and therefore no evidence yet exists that supports the hypothesis that tests potentiate learning of initially correctly retrieved items. To test this hypothesis, four between-subject conditions were used (see Figure 12). After initially studying the material, subjects took one or three cued recall tests. Half of the subjects in each of these conditions then restudied the material. A week later, all subjects took a final cued recall test. If tests potentiate learning of initially correct items, the effect of restudying on retention should have been enhanced in the three tests condition relative to the one test condition. This would be signified by an interaction between the test and restudy conditions such that the difference in retention between the restudy and no restudy conditions was larger for subjects who had taken more initial tests.

Because previous research has shown that restudy enhances retention more for items initially retrieved with low confidence than with high confidence (Butler et al., 2008), test-potentiated learning may also have a larger effect on low-confident correct items than high-confident correct items. All subjects made confidence ratings on the initial test to determine if test-potentiated learning varies over different levels of confidence. If the enhanced retention component is more prominent for low-confident items, the interaction between the test and restudy conditions should have been larger for low-confident items than for high-confident items. Logistic hierarchical linear modeling was used to analyze these results.

Figure 12. Design of Experiment 2. S = Study, T = Test, D = Distractor trial (Tetris game).

Experiment 2: The Enhanced Retention Component of Test-Potentiated Learning

The primary aim of Experiment 2 was to test the hypothesis that tests potentiate learning of initially correct items. This hypothesis has not been previously tested in part because in the past, feedback or restudy trials have often not been considered beneficial for items that have been correctly retrieved (Anderson et al., 1971; Kulhavy & Anderson, 1972; Pashler et al., 2005; Pashler, Rohrer et al., 2007). For instance, Pashler et al. (2005) found that feedback after a cued recall test did not improve later recall of initially correct items. Feedback had no effect on recall for correct items on either an immediate or a delayed (by 1 week) test. Further, it had no effect on recall for either high-confident or low-confident items. Findings like this one suggest that restudying an item after it has been learned cannot improve retention and no amount of prior testing can enhance an effect that does not exist.

However, more recent research has suggested that feedback after successful retrieval can enhance retention. Butler et al. (2008) had subjects answer general knowledge questions using a multiple-choice test and then rate their confidence in their answers. After half of the questions, subjects received feedback telling them which multiple-choice answer was correct. Subjects then took a final test either 5 min (Experiment 1) or 2 days (Experiment 2) later. Initially correct (and incorrect) items were more likely to be answered correctly on the final test when feedback was provided (see Figure 13). Further, this effect was modified by confidence. Low-confident correct answers benefited more from feedback than high-confident correct answers.

Experiment 2 extends Butler et al.'s (2008) findings to paired associate learning and cued recall tests. Further, it examines whether or not the number of tests prior to restudying modifies the effect. If tests potentiate learning of correctly retrieved items, restudying should have a

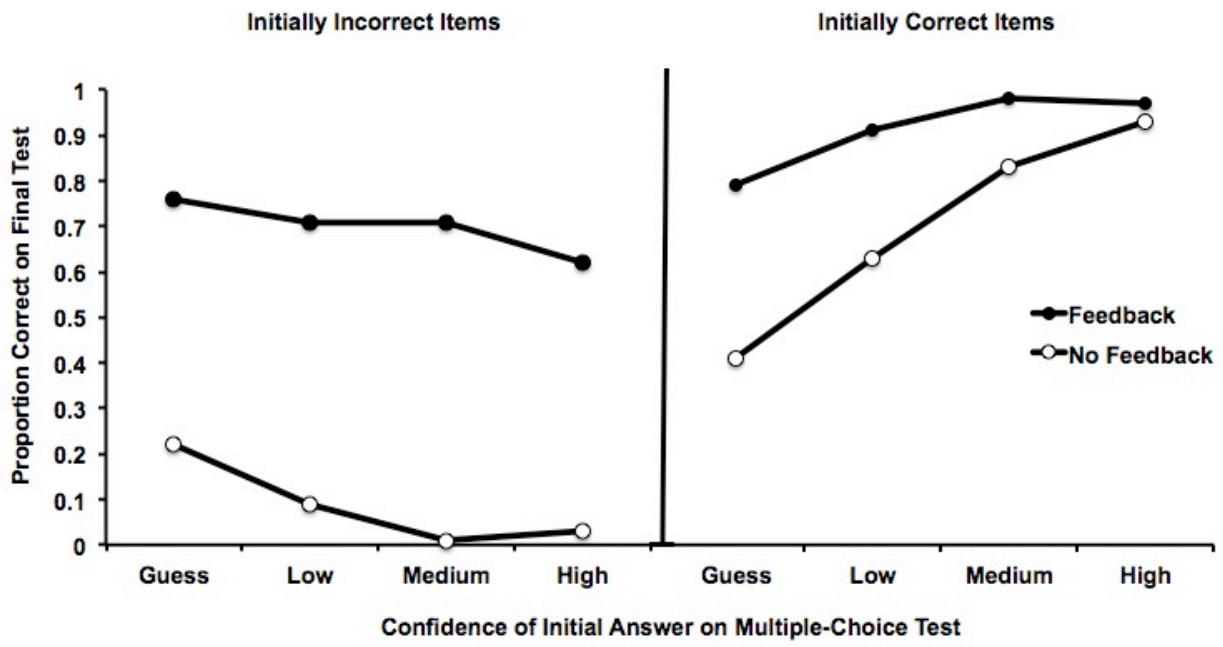

Figure 13. Mean proportion of items correct on the final test in Experiment 2 of Butler, Karpicke, and Roediger (2008) as a function of confidence on initial multiple-choice test and feedback condition for initially incorrect items (left panel) and initially correct items (right panel). Data estimated from their Figure 4.

greater benefit for correctly retrieved items after more tests have been taken. Given that restudying in general benefits low-confident correct items more than high-confident correct items, this potentiating effect may be more pronounced for low-confident correct items than high-confident correct items.

Method

Subjects

Seven hundred and forty-eight participants completed both Sessions 1 and 2 online through Amazon Mechanical Turk in exchange for $6. An additional 220 participants completed Session 1 but did not return to complete Session 2, creating a return rate of 77.3%. Of the participants who completed both sessions, 96 participants were excluded from the final analyses because on a post-experimental questionnaire they indicated they had written down words during the study phases. Additionally, 31 participants were excluded for not following directions. After these exclusions, 621 participants remained in the final analyses.

The remaining participants ranged in age from 18 to 69 (M = 31.3, SD = 10.5)[8]. More than half of the participants were female (n = 347). Education background varied widely from having less than a high school degree (n = 3) to having a doctorate (n = 11). Participation was limited to those residing in the United States (as determined by Amazon). Demographic characteristics did not vary across conditions (smallest p = .21).

Design

Like Experiment 1, this experiment had two parts. Part 1 was the same as in the previous experiment and was used as a baseline measure of memory ability.

[8] Due to a technical error, demographic information is incomplete for six participants. For each of these participants, information is missing for age, gender, and/or educational level.

Part 2 was the primary experiment and consisted of two sessions separated by 1 week (see Figure 12). The experiment used a 2 (tests: 1, 3) X 2 (restudy, no restudy) between-subjects factorial design. Participants were randomly assigned to one of four conditions: 3 tests and restudy (n = 143), 3 tests but no restudy (n = 175), 1 test and restudy (n = 155), and 1 test but no restudy (n = 148).

Materials

Stimuli used in the Part 1 baseline measure were the same as those used in the baseline measure in Experiment 1.

For Part 2, 60 weakly associated word pairs (e.g., *CITY – crowd)* were chosen from norms developed by Nelson, McEvoy, and Schreiber (2004; see Appendix B). All word pairs were unrelated to the words used in Part 1. Word pairs were chosen from a restricted range of forward strength association (range = 0.010-0.020, M = 0.013). Importantly, word pairs were chosen in groups of two [e.g., Pair A (*CITY – crowd*) and Pair B (*PEOPLE – world*)] such that cues from both Pair A (e.g., *CITY*) and Pair B (e.g., *PEOPLE*) were related to the target from Pair A (e.g., *crowd*). The forward strength association between the cue from Pair B and the target from Pair A (range = 0.021-0.320, M = 0.114) was always stronger than the forward strength association between the cue and target in the studied pairs. Pairs were chosen in this way to reduce subject's confidence in their answers.

For each study and test period, items were presented in a different random order determined on a participant-by-participant basis. For randomization purposes, each pair was treated individually. In this way, Pairs A and B from the same group could appear in any order with 0-58 intervening items between the related pairs. The post-experimental questionnaire given at the end of Session 1 was the same as the one used in Experiment 1. The post-experimental

questionnaire given at the end of Session 2 included only the questions pertaining to the participants' interest, perceived level of difficulty, and effort.

Procedure

As in Experiment 1, participants first answered a series of demographic questions before beginning the experiment. They were told this experiment would take place over two sessions with Session 2 taking place 1 week after Session 1. Session 1 consisted of two parts: in Part 1 participants learned a series of words and in Part 2 they learned a series of word pairs. Instructions were the same for all participants.

The baseline measure (Part 1 of Session 1) was identical to the baseline measure of the previous experiment. In Part 2 of Session 1 (the main experiment), participants first studied 60 word pairs. Each pair was presented individually for 3 s with a 500 ms interstimulus interval. This presentation rate was chosen because pilot work indicated it produced more low confident correct items than a longer presentation rate of 5 s. Participants were not told about the relationship between the grouped pairs. After studying all of the word pairs, participants were given a distractor task consisting of five addition and/or subtraction problems (5 s to answer each problem, 500 ms interstimulus interval). Next, participants were given a self-paced cued recall test on the word pairs. Each cue was presented individually on the screen, and participants were instructed to type the matching target word. After typing their answer, participants indicated how confident they were in their answer using a scale ranging from 0-100, with 0 indicating their answer was a guess and 100 indicating complete confidence their answer was correct. At the beginning of each trial, a marker was positioned in the middle of a line at a position corresponding to 50% confidence. They reported their confidence level by sliding the marker on the line, which was labeled with multiples of 10, to the appropriate number with their mouse.

After moving the marker to the number corresponding with their level of confidence, participants pushed a button on the screen using their mouse to move on to the next trial.

After completing all 60 test trials, the procedure varied by condition. Participants assigned to the 1 test but no restudy condition were finished with Session 1 and filled out a post-experimental questionnaire immediately after the initial test. Participants in the 1 test and restudy condition played two games of Tetris, each lasting 5 min and 30 s. The Tetris games served as distractor tasks to equate the spacing between the initial study and restudy phases in the 1 test and 3 test conditions. The remaining participants were given two additional cued recall tests. These additional tests were not self-paced, and participants did not give confidence ratings. Instead, each cue was presented individually on the screen for 5 s (500 ms interstimulus interval) during which time participants attempted to type the corresponding target. After taking the additional tests, participants in the 3 tests but no restudy condition were finished with Session 1 and were given a post-experimental questionnaire. Participants assigned to both restudy conditions restudied all 60 word pairs after playing Tetris or taking the additional tests. The restudy phase proceeded in the same fashion as the initial study phase. After restudying, participants were given a post-experimental questionnaire.

One week after completing Session 1, participants completed Session 2, which consisted of a final self-paced cued recall test. Each cue word was presented individually, and participants were instructed to type the corresponding target word for each cue. After typing their response, they pressed a button on the screen using their mouse to continue to the next trial. After being tested on all 60 pairs, participants completed another post-experimental questionnaire.

Results

Results from the baseline measure from Part 1 are presented first. To preview, as in Experiment 1 no group differences were found on this measure, and therefore it was not included in subsequent analyses.

Results from the main experiment (Part 2) are presented next. The results are divided into four sections: initial test results (including all tests given during Session 1), final test results (unconditionalized), conditional analyses, and accuracy of the confidence ratings.

Finally, results from the post-experimental questionnaires given after both Session 1 and Session 2 are presented.

Baseline Measure

As in Experiment 1, the range in the proportion of words recalled was large (min = .00, max = 1.00), but recall did not vary by condition, $F < 1$. Participants in the 3 tests and restudy (M = .41), 3 tests but no restudy (M = .40), 1 test and restudy (M = .39), and 1 test but no restudy (M = .41) conditions all recalled approximately the same proportion of words, indicating that there were no pre-experimental group differences in memory ability. Further, there were no differences in the number of critical items (i.e., items not presented at study but related to all items in one of the three lists) recalled as false alarms, $F(3, 624) = 1.83$, $p = .14$, $\eta_p^2 = .009$. Participants in all conditions recalled about one out of the three (1/3) critical items: 3 tests and restudy (M = .31), 3 tests but no restudy (M = .27), 1 test and restudy (M = .30), and 1 test but no restudy (M = .35). This result provides further evidence that there were no pre-experimental group differences. These findings are particularly important given the week delay and subsequent loss of participants who did not return for Session 2. These results suggest that the reduction in participants did not differentially affect any condition.

Main Experiment

Initial test. After the initial study phase, all participants took an initial test. A one-way ANOVA indicated that there were no differences between conditions, $F(3, 617) = 1.01$, $p = .39$, $\eta_p^2 = .005$, which was expected given that the experimental manipulation had not yet occurred. Participants in the 3 tests and restudy ($M = .47$), 3 tests but no restudy ($M = .45$), 1 test and restudy ($M = .43$), and 1 test but no restudy ($M = .46$) conditions all recalled approximately the same proportion of word pairs on the initial test.

Half of the participants took two additional tests[9]. Recall varied across these three tests, $F(2, 630) = 47.57$, $p < .001$, $\eta_p^2 = .13$. More words were recalled on the initial test ($M = .46$) than on both the second ($M = .43$) and third ($M = .44$) tests, $t(316) = 8.54$, $p < .001$, $d = .14$ and $t(317) = 6.33$, $p < .001$, $d = .10$, respectively. Further, more words were recalled on the third test than on the second test, $t(316) = 3.56$, $p < .001$, $d = .04$, although this difference was small as indicated by the small effect size. Collapsing across these three tests, there was no difference between participants in the restudy condition ($M = .45$) and those in the no restudy condition ($M = .44$), $F < 1$, as was expected given that the restudy phase came after these tests. Additionally, there was no interaction between restudy condition and test number, $F < 1$.

Final test. As in Experiment 1, the final test data were analyzed multiple ways. First, the overall proportions of items recalled on the final test were compared across conditions. In the following section, recall data were analyzed by conditionalizing the data on initial test performance. Further, the conditional data were then analyzed as a function of confidence ratings made during the initial test. This latter analysis is of primary interest because it was used to determine if the effects of prior tests and restudying were modified by confidence.

[9] Due to a technical error, data from the second test is missing for one participant.

Overall, recall on the final test was greater for participants who had taken three prior tests (M = .37) than those who had taken one prior test (M = .24; see Figure 14), $F(1, 617) = 67.92$, $p < .001$, $\eta_p^2 = .10$. Recall was also greater for participants who had restudied the items (M = .33) relative to those who had not (M = .27), $F(1, 617) = 16.89$, $p < .001$, $\eta_p^2 = .03$. Further, there was no interaction between number of prior tests and restudy condition, $F(1,617) = 2.71$, $p = .10$, $\eta_p^2 = .004$, indicating there was no overall potentiating effect of the prior tests. However, as can be seen in Figure 14, the pattern of results resembled a test-potentiation-like pattern. Although not significant, numerically there was a larger difference between the restudy and no restudy conditions in the 3 Test condition (M_{diff} = .09) than in the 1 Test condition (M_{diff} = .04).

Conditional analyses. Final test data were analyzed as a function of initial test performance. First, the proportion of initially correct items that were recalled on the final test (i.e., retained over the week delay) was examined (see Figure 15). Participants who took three tests (M = .67) retained more items than those who took one test (M = .44), $F(1, 617) = 192.63$, $p < .001$, $\eta_p^2 = .24$. Additionally, participants who restudied the material (M = .59) retained more items than those who did not (M = .53), $F(1, 617) = 13.50$, $p < .001$, $\eta_p^2 = .02$, suggesting that the restudy trial was beneficial for correctly recalled items. That is, subjects accrued information during restudy related to correctly recalled items, benefiting subsequent memory. This result is consistent with Butler et al.'s (2008) findings using multiple choice tests, but is contradictory to Pashler et al.'s earlier (2005) finding that feedback was not beneficial for correctly retrieved items in paired-associate learning. There are a number of methodological differences between the present experiment and Pashler et al.'s experiment that could explain this discrepancy. For example, whereas the present experiment used low-associate word pairs, Pashler et al. used

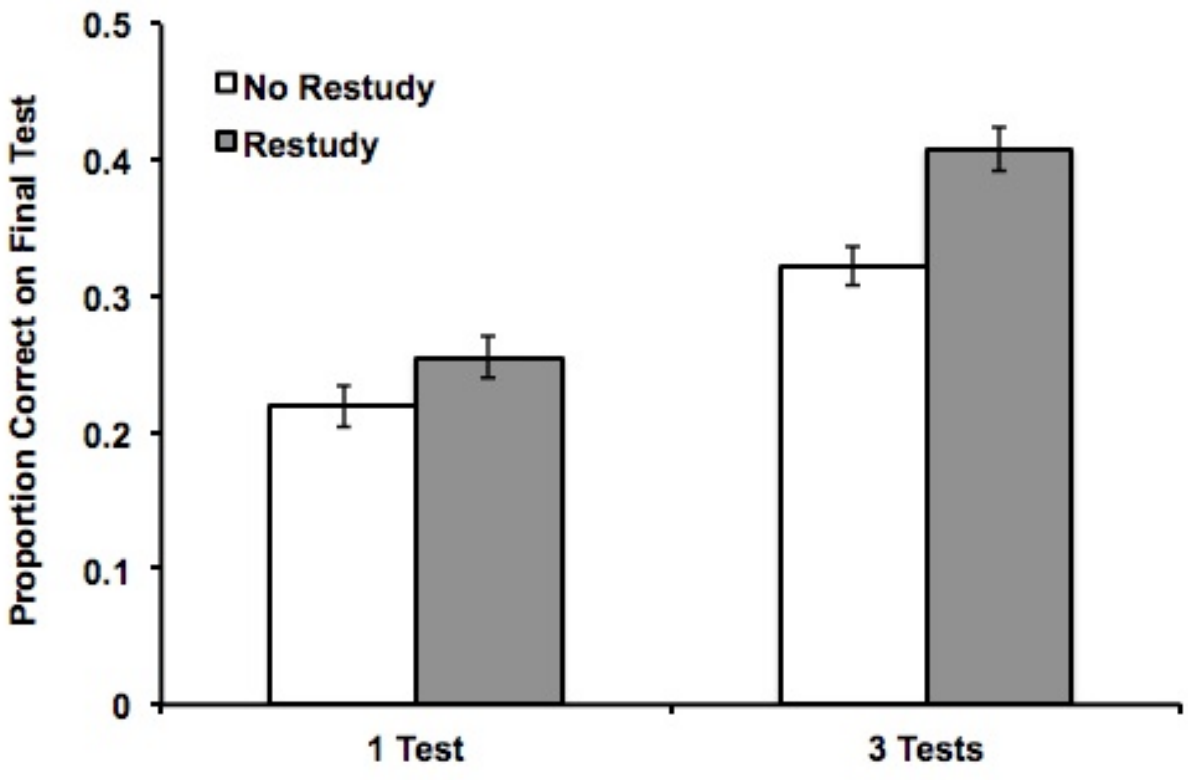

Figure 14. Mean proportion of items correct on the final test in Experiment 2 as a function of test and restudy conditions. Error bars represent standard errors of the mean.

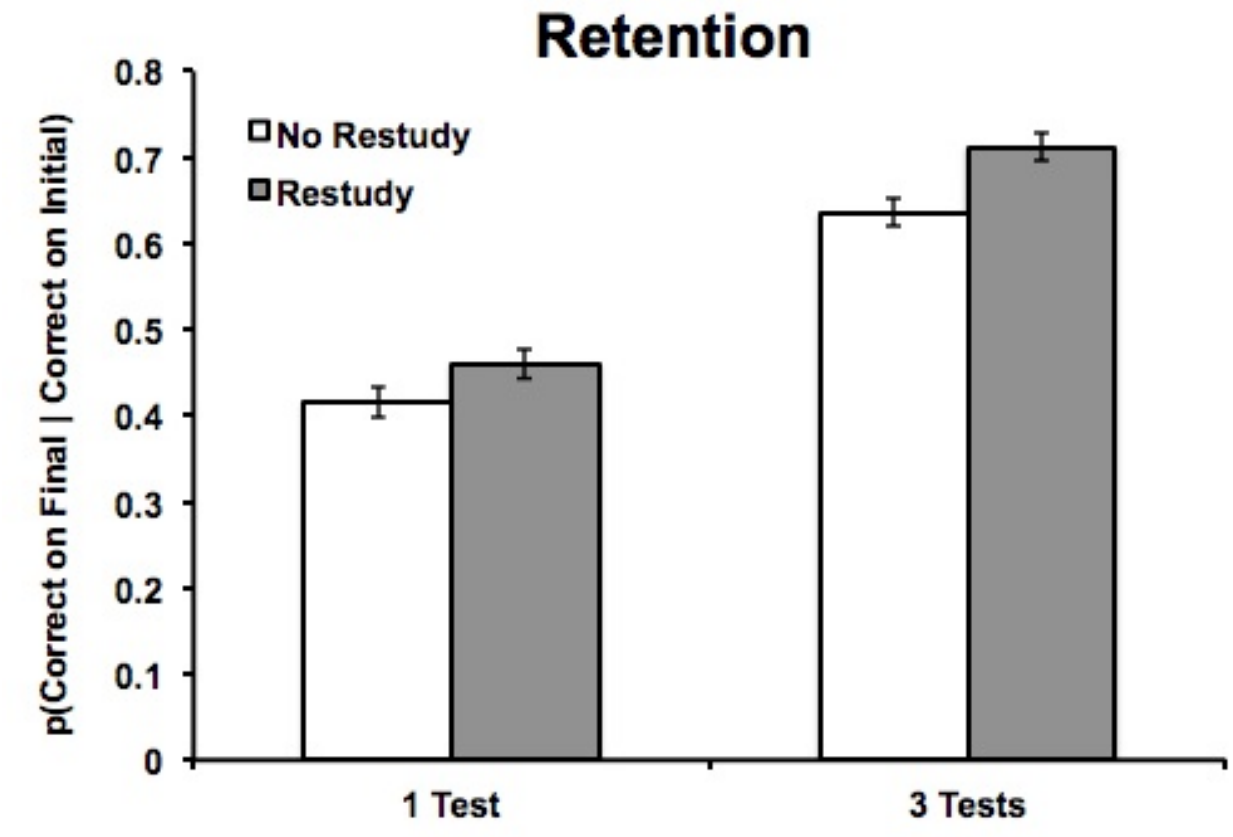

Figure 15. Mean proportion of items retained from the initial test to the final test in Experiment 2 as a function of test and restudy conditions. Error bars represent standard errors of the mean.

Luganda-English word pairs, which were essentially unrelated pairs to the subjects. Further, in the present experiment, the materials were chosen in such a way as to encourage low-confident correct items, which were hypothesized to benefit the most from feedback (see Methods section). The foreign language-English word pairs used in the Pashler et al. experiment may not have produced as many low-confident correct items. In addition, in the Pashler et al. experiment, feedback was given immediately after the test whereas as in the present experiment restudy was delayed by several minutes and intervening items. Delaying feedback may have enhanced the effect of restudy (e.g., Butler, Karpicke, & Roediger, 2007).

Did the number of prior tests modify the effect of restudy on initially correct items? In other words, did the prior tests potentiate learning of correctly recalled items? An interaction between number of prior tests and restudy condition would have suggested a potentiating effect, but there was no interaction present, $F < 1$. This result indicates that the additional tests may not have potentiated learning of initially correct items.

However, as can be seen in Figure 16, confidence ratings for initially correct items were highly negatively skewed (skewness = -2.07) meaning many more initially correct items were given a high confidence rating than a low confidence rating. The potentiating benefit of the additional tests may be larger (or only present) for items given lower confidence ratings, and therefore the negative skew may be masking a potentiating effect. To determine how confidence may interact with the effect of test-potentiated learning, recall of initially correct items was analyzed as a function of confidence.

To assess the effect of confidence on final test recall for initially correct items, a two-level logistic hierarchical linear regression analysis was conducted (Raudenbush & Bryk, 2002).

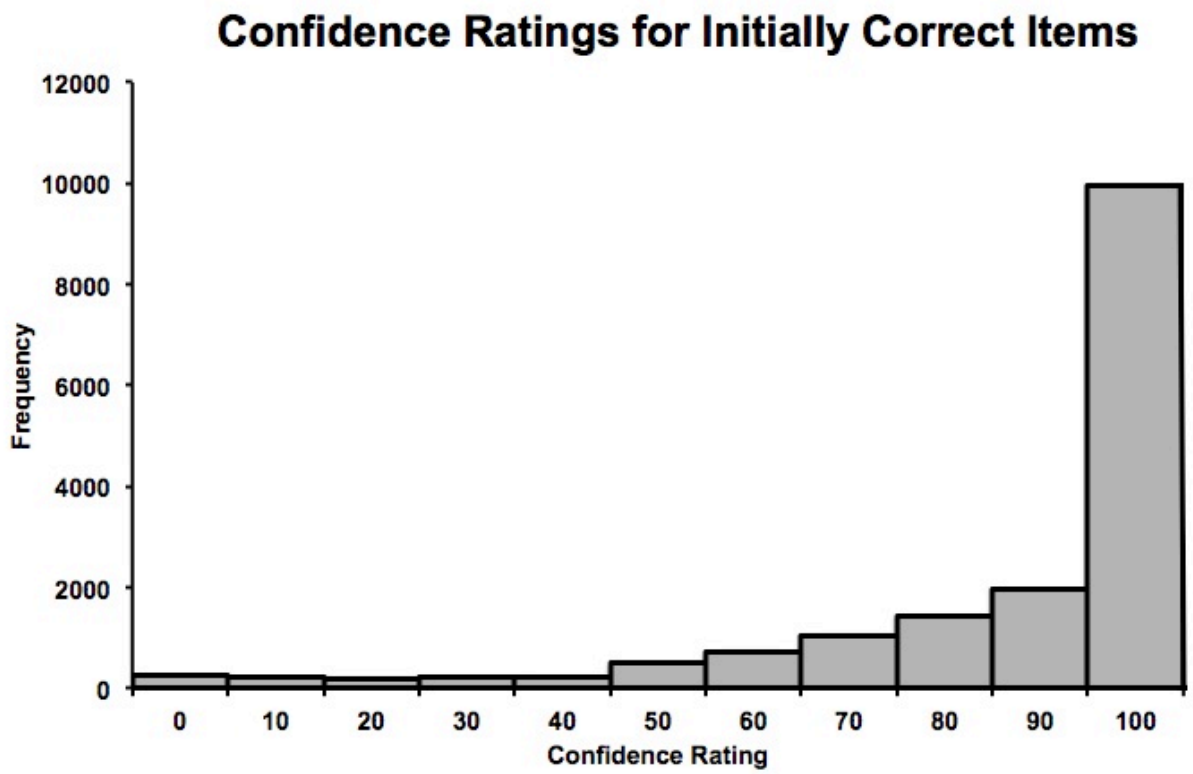

Figure 16. Frequency of all initially correct items across all participants given each confidence rating.

A multilevel regression analysis was used because of the nested structure of the data; items were nested within subjects, and therefore the probability that a given item was recalled on the final test was not independent of other items from the same subject. A multilevel regression model takes this into account by allowing the intercept and/or slope to vary across subjects. A logistic regression analysis was used because the outcome at level 1 was binomial. That is, level 1 modeled the probability that a given item was recalled on the final test, and for any given item there were only two possible outcomes: recalled (1) or not recalled (0).

The tested model is presented in Figure 17. At level 1 (the item level), confidence rating on the initial test was the only predictor. Confidence ratings were centered within subjects so that a rating of 0 was equal to a given subject's mean confidence rating for initially correct items. At level 2 (the subject level), predictors for both the slope and intercept included test condition (dummy coded: 0 = 1 test, 1 = 3 tests), restudy condition (dummy coded: 0 = no restudy, 1 = restudy), an interaction between test and restudy conditions, and proportion of items recalled on the initial test (grand mean centered so that 0 = average proportion of recalled items). The proportion of items recalled on the initial test was included as a measure of individual difference in performance to reduce variance and more accurately model the data.

The model was compared to an intercepts-only model to determine if the predictors as a group accounted for a significant proportion of the variance (Raudenbush & Bryk, 2002). The full model was significantly different than the intercepts-only model, $\chi^2(11) = 549.44$, $p < .001$, indicating that together the predictors accounted for more of the variance than a model that only considered variability in subjects. Next, a model in which confidence (level 1 slope) was modeled as a nonrandomly varying effect was compared to one in which confidence was

Level 1 (item)

Y_{ij} = p(item *i* was recalled on the final test by person *j*)

$\ln[Y_{ij}/(1-Y_{ij})] = \beta_{0j} + \beta_{1j}$*(confidence)

Level 2 (person)

$\beta_{0j} = \gamma_{00} + \gamma_{01}$(test condition) + γ_{02}(restudy condition) + γ_{03}(initial test) + γ_{04}(test condition * restudy condition) + μ_{0j}

$\beta_{1j} = \gamma_{10} + \gamma_{11}$(test condition) + γ_{12}(restudy condition) + γ_{13}(initial test) + γ_{14}(test condition * restudy condition) + μ_{1j}

Figure 17. The logistic hierarchical linear regression model used to analyze the final recall data for both initially correct and initially incorrect items in Experiment 2.

modeled as a randomly varying effect. The two models were significantly different, $\chi^2(2) = 8.27$, $p = .02$, indicating that modeling slope as a randomly varying effect accounted for more of the variance. This suggests that there are individual differences in the relationship between confidence and final test accuracy. For this reason, slope was modeled as a randomly varying effect.

Results from the analysis are listed in Table 2, and the data are presented in Figures 18 and 19. Figure 18 presents the proportion of initially correct items recalled on the final test in each condition as a function of the raw confidence rating (i.e., the value subjects indicated represented their level of confidence). Figure 19 presents the same information as a function of confidence ratings that have been centered on each subjects mean, as was done in the regression model. As expected, the effects on the level 1 intercept closely mirror the results obtained from the earlier discussed ANOVA. Participants who took three tests retained more items than those who took one test, $\gamma_{01} = 1.12$, $Z = 10.09$, $p < .001$. Participants who restudied retained more items than those who did not restudy, $\gamma_{02} = 0.31$, $Z = 2.68$, $p = .01$. Further, there was no test condition by restudy condition interaction, $\gamma_{04} = 0.10$, $Z = 0.61$, $p = 0.54$, once again indicating that when collapsing across confidence ratings, there was no overall potentiating effect of tests for initially correct items. Finally, there was an effect of initial test performance such that participants who recalled more items on the initial test retained a larger proportion of those items on the final test, $\gamma_{03} = 1.69$, $Z = 8.38$, $p < .001$. The level 1 slope represents the effect of confidence on final recall. As can be seen in Figures 18 and 19, there was a main effect of confidence, $\gamma_{10} = 0.01$, $Z = 7.62$, $p < .001$. That is, for each 10-point increase on the confidence rating scale, the odds that a given item would be recalled on the final test increased by an

Table 2

Results from the two-level hierarchical logistic regression model for initially correct items.

Symbol		Effect	Coefficient	Standard Error	*z* value	*p* value
For β_{0j}, intercept	γ_{00}	Intercept	-0.47	0.082	-5.69	< 0.001
	γ_{01}	Test condition	1.12	0.111	10.09	< 0.001
	γ_{02}	Restudy condition	0.31	0.114	2.68	0.01
	γ_{03}	Initial test	1.69	0.202	8.38	< 0.001
	γ_{04}	Test X Restudy	0.10	0.160	0.61	0.54
For β_{1j}, slope	γ_{10}	Confidence	0.01	0.002	7.62	< 0.001
	γ_{11}	Confidence X Test	0.003	0.002	1.06	0.29
	γ_{12}	Confidence X Restudy	-0.01	0.002	-2.32	0.02
	γ_{13}	Confidence X Initial test	0.004	0.005	0.80	0.42
	γ_{14}	Confidence X Test X Restudy	0.004	0.003	1.07	0.28

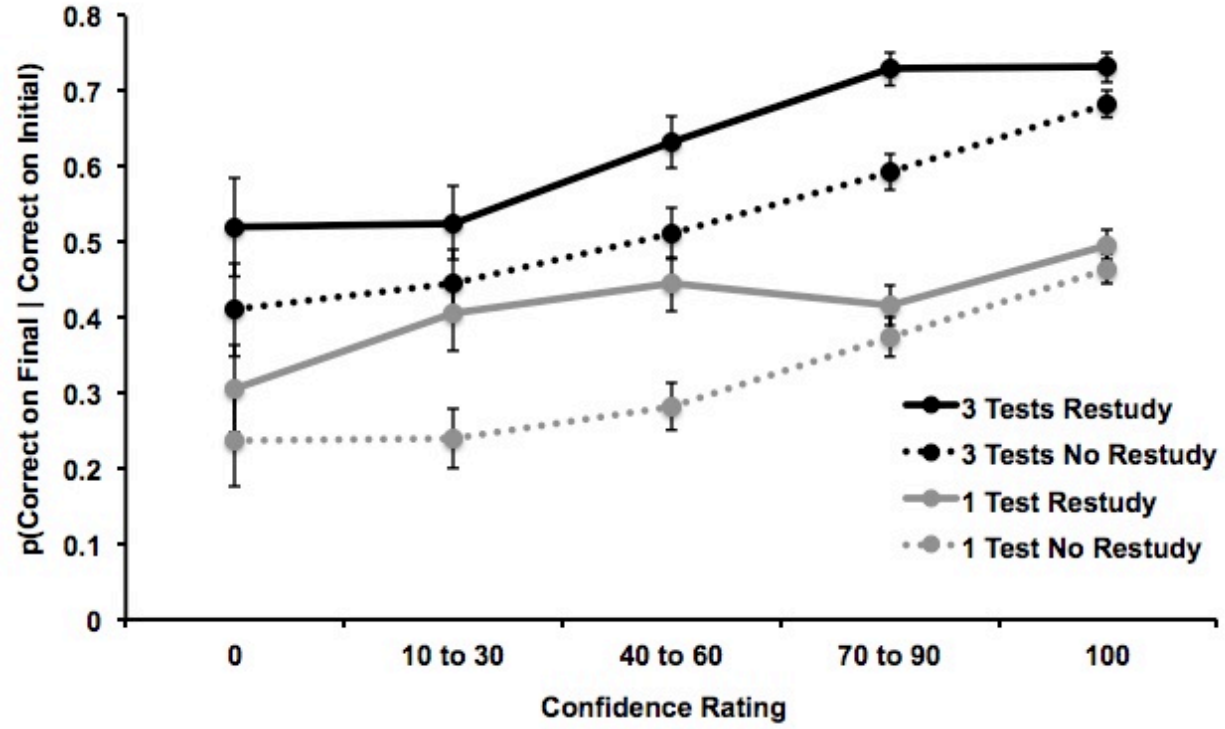

Figure 18. Mean proportion of initially correct items that were recalled on the final test in Experiment 2 as a function of initial confidence ratings and learning condition. Error bars represent standard errors of the mean.

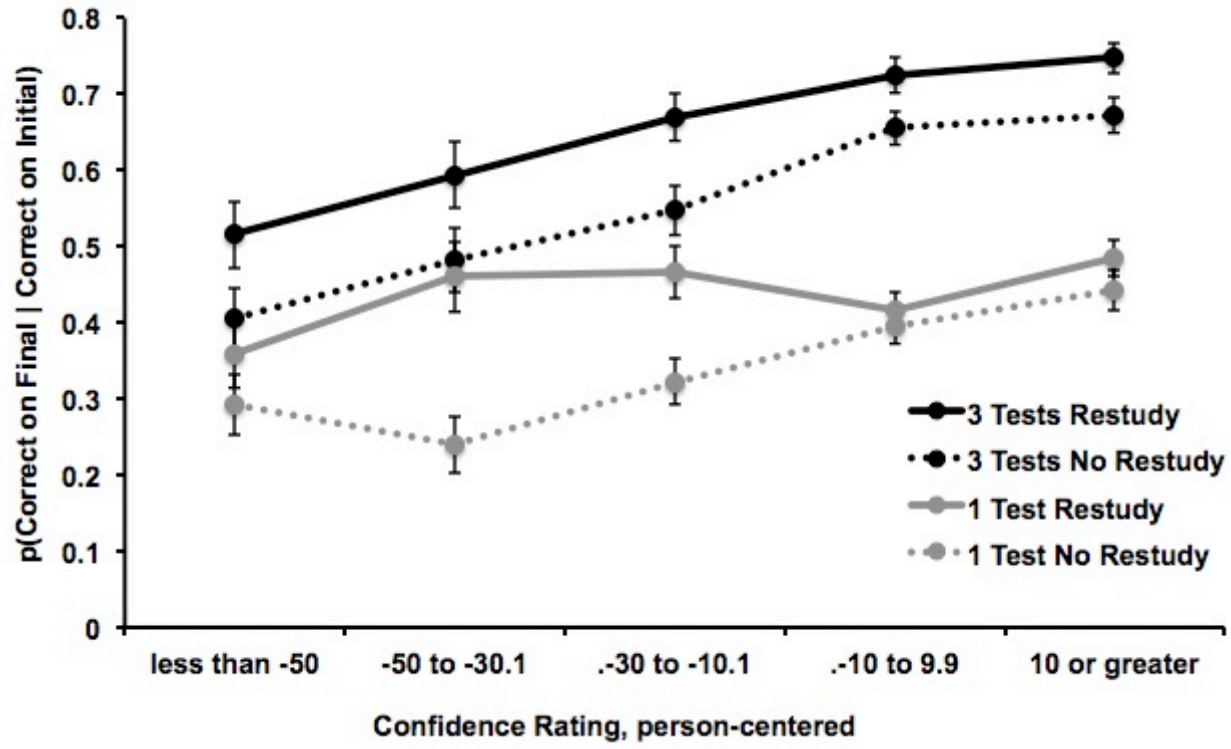

Figure 19. Mean proportion of initially correct items that were recalled on the final test in Experiment 2 as a function of initial confidence ratings and learning condition. Confidence ratings were person-centered such that 0 represents the mean confidence rating of initially correct items for a given subject. A positive rating indicates an above-average confidence rating for a given subject. A negative rating indicates a below-average confidence rating for a given subject. Error bars represent standard errors of the mean.

average of 14.3%. However, this effect was modified by an interaction. The effect of confidence ratings depended on whether or not subjects were given a restudy opportunity, as shown by a significant confidence by restudy interaction, $\gamma_{12} = -0.01$, $Z = -2.32$, $p = .02$. Because the restudy variable was dummy coded such that a value of 0 indicated no restudy opportunity, the above described increase in odds (14.3%) for each 10-point increase in confidence was for subjects who did not restudy. For subjects who did restudy, a 10-point increase in confidence ratings only increased the odds an item would be recalled by 8.5%. Alternatively, the interaction could be interpreted in the reverse direction. That is, as confidence ratings increased, the enhancing effect of restudying decreased. That is, having a chance to restudy the material benefited items that were correctly recalled with low confidence more than items correctly recalled with high confidence. This pattern replicates results found by Butler et al. (2008).

Because restudying benefited low-confident correct items more than high confident correct items, test-potentiated learning may only be present for items initially recalled with low confidence. That is, testing may enhance the benefit of restudying low-confident items but not enhance (or not enhance to the same degree) the benefit of restudying high confident items. This would be indicated by a three-way interaction between test condition, restudy condition, and confidence ratings. However, this pattern did not emerge. The three-way interaction was not significant, $\gamma_{14} = 0.004$, $Z = 1.07$, $p = .28$, indicating that the effect of testing on subsequent encoding did not vary across confidence ratings. This result, along with the non-significant two-way interaction between test condition and restudy condition suggests that tests may not potentiate learning for items that are initially correctly recalled.

In addition to the analyses conducted on initially correct items, several analyses were also conducted on items that were not initially recalled. Given the one week delay between Session 1

and the final test, the proportion of initially incorrect items that were retrieved on the final test was near floor in all conditions (see Figure 20). For this reason, interpretation of differences (or lack thereof) between conditions is difficult, but the results are presented here for completeness.

First, final test data for initially incorrect items were analyzed without taking into account confidence ratings. As can be seen in Figure 20, participants who took three tests (M = .10) retrieved a larger proportion of initially incorrect items than those who took only one test (M = .07), $F(1, 617) = 5.54$, $p = .02$, $\eta_p^2 = .009$. Participants who restudied the material (M = .12) recalled a larger proportion of initially incorrect items than those who did not restudy (M = .04), $F(1, 617) = 81.42$, $p < .001$, $\eta_p^2 = .11$. The effect of restudying was numerically larger for participants who took three intervening tests (M_{diff} = .10) than for participants who took only one intervening test (M_{diff} = .07), but this difference was only marginally significant, $F(1, 617) = 3.75$, $p = .053$, $\eta_p^2 = .006$. Although this interaction was only marginally significant, the pattern suggests that prior tests may have potentiated learning for initially incorrect items. This pattern was further explored using two follow-up t-tests with a Bonferroni correction (i.e., significance indicated at $p < .025$). For subjects who restudied, subjects who took three intervening tests (M = .14) recalled more initially incorrect items on the final test than subjects who took only one intervening test (M = .10), $t(296) = 2.73$, $p = .007$, $d = .32$. In contrast, for subjects who did not restudy, final recall of initially incorrect items did not differ between test conditions ($M_{3\ Tests}$ = .05 vs. $M_{1\ Test}$ = .04), $t(321) = 1.08$, $p = .28$, $d = .12$. These results further indicate that the prior tests potentiated learning of initially incorrect items and that the effects of this potentiation lasted over the week delay. However, given the floor effects, interpretation of these results must be made cautiously.

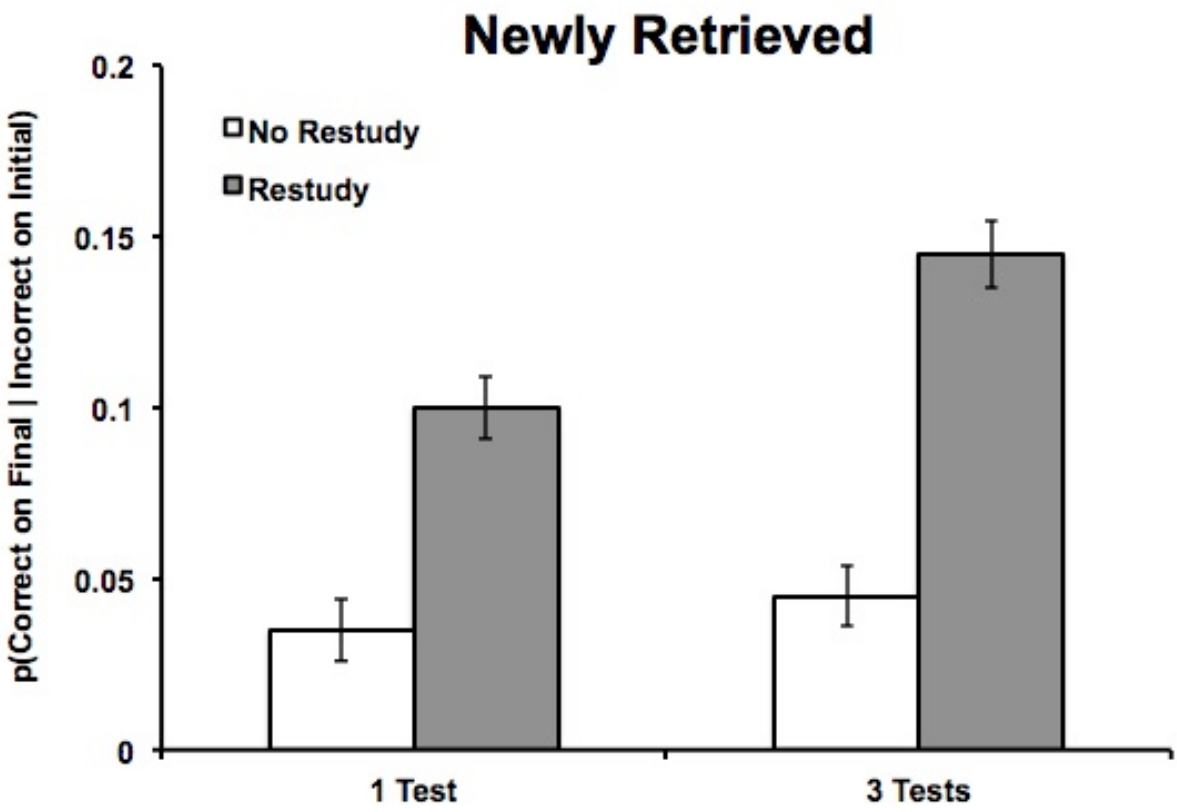

Figure 20. Mean proportion of items newly retrieved on the final test in Experiment 2 as a function of test and restudy conditions. Error bars represent standard errors of the mean.

Next, these data were analyzed as a function of confidence ratings using a two-level logistic hierarchical regression analysis. The same model (Figure 17) used for correctly recalled items was also used for incorrectly recalled items. The full model was significantly different than an intercepts-only model, $\chi^2(11) = 302.36$, $p < .001$, indicating that the predictors as a group accounted for a significant proportion of the variance. Further, a randomly varying effect slope model was significantly different than a nonrandomly varying effect slope model, $\chi^2(2) = 8.09$, $p = .02$, so the randomly varying effect slope model was used in the following analysis.

The results from the analysis are presented in Table 3, and the data are presented in Figures 21 (raw confidence ratings) and 22 (person-centered confidence ratings). Looking first at effects on the level 1 intercept, this analysis indicates that participants who took three intervening tests recalled a larger proportion of initially incorrect items on the final test than those who took only one intervening test, $\gamma_{01} = 0.38$, $Z = 2.16$, $p = .03$. Similarly, the analysis indicates that participants who restudied recalled a larger proportion of these items than those who did not restudy, $\gamma_{02} = 1.43$, $Z = 8.46$, $p < .001$. These main effects are consistent with the findings from the earlier ANOVA analysis. However, unlike in the ANOVA analysis, the logistic hierarchical linear regression model indicated that there was no interaction between test condition and restudy condition, $\gamma_{04} = -0.03$, $Z = -0.15$, $p = .88$, suggesting that tests did not potentiate learning for initially incorrect items. In addition, the regression analysis also indicated a significant effect of initial recall, $\gamma_{03} = 3.03$, $Z = 10.83$, $p < .001$. Participants who recalled more items on the initial test recalled a larger proportion of initially incorrect items on the final test.

This analysis also examined the relationship between confidence (level 1 slope) and final recall. These results indicated that subjects were more likely to later recall the correct target

Table 3

Results from the two-level hierarchical logistic regression model for initially incorrect items.

Symbol		Effect	Coefficient	Standard Error	*z* value	*p* value
For β_{0j}, intercept	γ_{00}	Intercept	-4.05	0.137	-29.53	< 0.001
	γ_{01}	Test condition	0.38	0.178	2.16	0.03
	γ_{02}	Restudy condition	1.43	0.169	8.46	< 0.001
	γ_{03}	Initial test	3.03	0.280	10.83	< 0.001
	γ_{04}	Test X Restudy	-0.03	0.226	-0.15	0.88
For β_{1j}, slope	γ_{10}	Confidence	0.01	0.004	2.50	0.01
	γ_{11}	Confidence X Test	-0.004	0.005	-0.79	0.43
	γ_{12}	Confidence X Restudy	-0.003	0.004	-0.72	0.47
	γ_{13}	Confidence X Initial test	-0.002	0.007	-0.35	0.73
	γ_{14}	Confidence X Test X Restudy	0.01	0.006	1.21	0.23

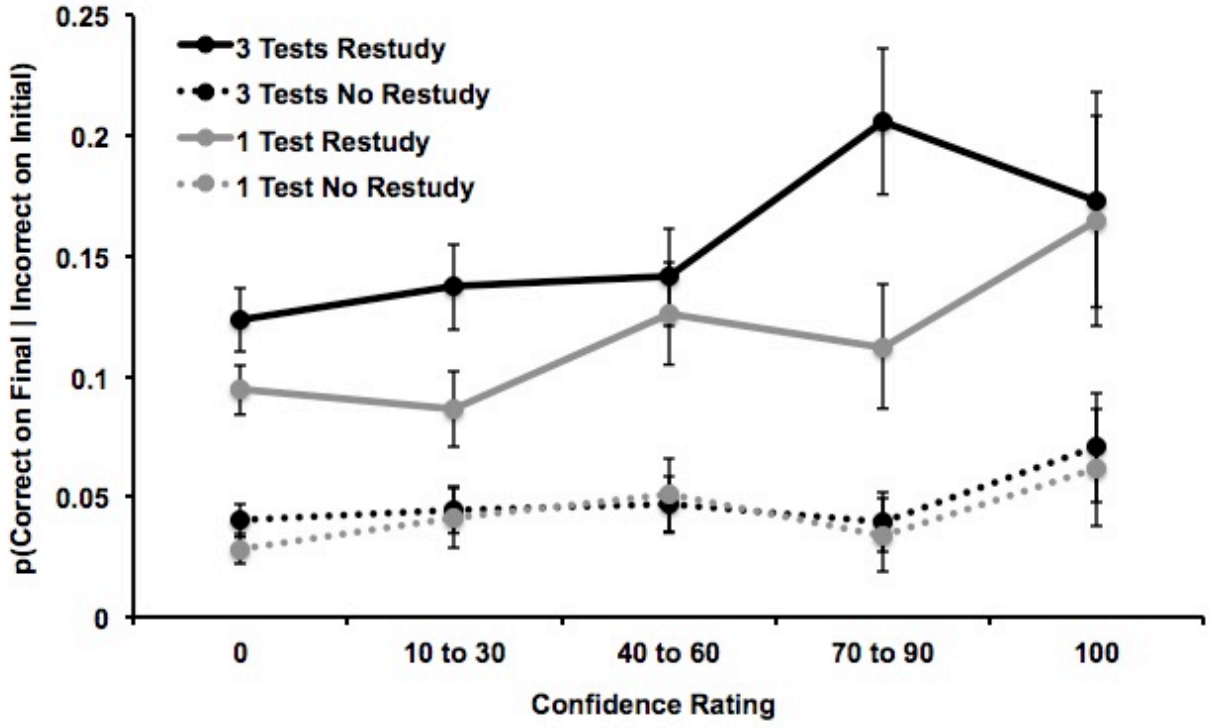

Figure 21. Mean proportion of initially incorrect items that were recalled on the final test in Experiment 2 as a function of initial confidence ratings and learning condition. Error bars represent standard errors of the mean.

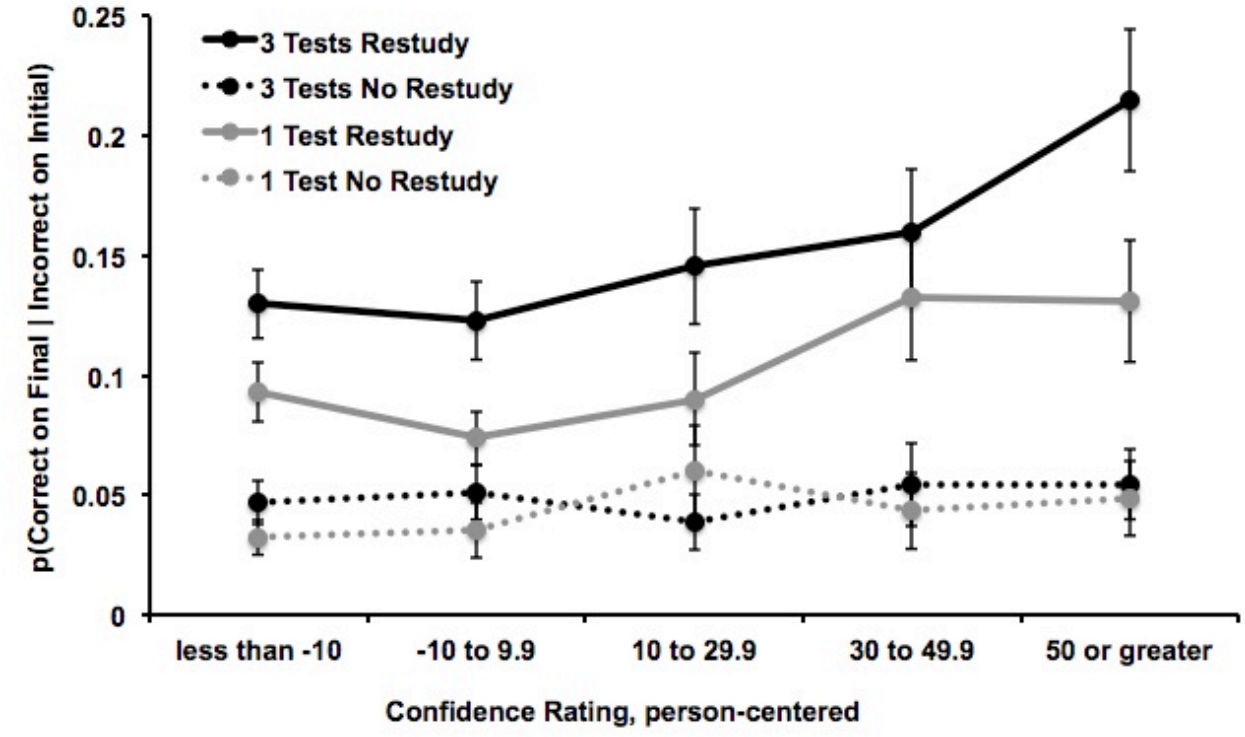

Figure 22. Mean proportion of initially incorrect items that were recalled on the final test in Experiment 2 as a function of initial confidence ratings and learning condition. Confidence ratings were person-centered such that 0 represents the mean confidence rating of initially incorrect items for a given subject. A positive rating indicates an above-average confidence rating for a given subject. A negative rating indicates a below-average confidence rating for a given subject. Error bars represent standard errors of the mean.

when the initially incorrect response was made with higher confidence, $\gamma_{10} = 0.01$, $Z = 2.50$, $p = .01$. That is, for every 10-point increase in confidence rating, the odds a given item would be recalled on the final test increased by 8.99%. No other predictors had an effect on the level 1 slope indicating that test condition, restudy condition, and level of initial recall had no effect on the relationship between confidence and final recall.

The significant relationship between confidence and later recall is inconsistent with a previous finding by Butler et al. (2008), which indicated that for initially incorrect items, confidence did not predict recall on a delayed test (see Figure 13). However, it is consistent with previous research on the hypercorrection effect (Butterfield & Metcalfe, 2001, 2006), which indicates that when feedback is provided incorrect responses made with high confidence are more likely to be corrected than those made with low confidence. Although Butler et al. found that a hypercorrection effect went away and may have even reversed over long retention intervals, other research has shown that the hypercorrection effect can persist over retention intervals of at least a week (e.g., Butler, Fazio, & Marsh, 2011; Kulhavy, Yekovich, & Dyer, 1976). Because a hypercorrection effect is found after feedback, one may have expected that the relationship between confidence and final recall would be present only in the restudy condition. On visual inspection, the pattern shown in Figures 21 and 22 appears to be somewhat consistent with this hypothesis. However, restudy condition did not interact with confidence suggesting that this effect was not dependent on subjects having a restudy opportunity. However, this result may have been due to the floor effects in the no restudy condition. Further research is needed to determine how restudying may effect the relationship between confidence and later recall for initially incorrect items.

Accuracy of confidence ratings. Finally, the accuracy of the confidence ratings was examined in two ways. First, the calibration (i.e., absolute accuracy, or the degree to which the overall level of confidence corresponded with the actual level of overall recall) was examined. Next, the resolution (i.e., relative accuracy, or the degree to which confidence ratings corresponded with recall of items relative to other items) was examined (see Koriat & Goldsmith, 1996; Nelson, 1984; Nelson & Dunlosky, 1991).

A calibration analysis indicated that participants were generally overconfident in their predictions of their future recall. That is, a 2 (confidence rating, initial test performance) X 4 (condition) mixed ANOVA indicated that the average confidence rating (M = .50) was greater than the average initial test recall (M = .45), $F(1, 617) = 109.80$, $p < .001$, $\eta_p^2 = .15$. As expected given that the experimental manipulation had not yet occurred, this overconfidence pattern was observed in every condition (3 tests and restudy: $M_{Confidence} = .50$, $M_{Recall} = .47$; 3 tests but no restudy: $M_{Confidence} = .50$, $M_{Recall} = .45$; 1 test and restudy: $M_{Confidence} = .47$, $M_{Recall} = .43$; 1 test but no restudy: $M_{Confidence} = .51$, $M_{Recall} = .46$). Confirming this finding, there was no significant effect of condition or interaction, $F(3, 617) = 1.18$, $p = .32$, $\eta_p^2 = .006$ and $F < 1$, respectively.

Next, resolution was analyzed in two ways: the relative correspondence between confidence and initial test recall and the relative correspondence between confidence and final test recall. The former comparison determines if relative correspondence was initially equated across groups, as is expected given that the experimental manipulations had not yet occurred. The latter comparison determines how additional tests and restudy trials affected this relationship. Resolution was computed using Goodman-Kruskal gamma correlations between confidence and performance.

Resolution between confidence and initial test performance was near ceiling and, as expected, the gamma correlations were equated across conditions (3 tests and restudy: $M = .92$; 3 tests but no restudy: $M = .92$; 1 test and restudy: $M = .93$; 1 test but no restudy: $M = .92$; see Figure 23). This was confirmed using a 2 X 2 ANOVA, which indicated there was no main effect of number of tests or restudy condition, $F(1,617) = 2.06$, $p = .15$, $\eta_p^2 = .003$ and $F < 1$, respectively. Further, there was no test by restudy condition interaction, $F(1,617) = 1.33$, $p = .25$, $\eta_p^2 = .002$. In contrast, resolution between confidence and final test performance did vary across conditions[10]. Gamma correlations were larger for participants who took three tests ($M = .72$) than for those who took one test ($M = .64$), $F(1, 600) = 18.33$, $p < .001$, $\eta_p^2 = .03$. Additionally, gamma correlations were larger for participants who did not restudy ($M = .73$) than for those who did restudy ($M = .63$), $F(1, 600) = 23.95$, $p < .001$, $\eta_p^2 = .04$. There was no test by restudy interaction, $F < 1$.

The effects of prior tests and restudying on resolution may at first seem contradictory. Although both testing and restudying enhanced recall probability on the final test, testing increased resolution whereas restudying decreased resolution. The reason for these opposing effects is likely due to the different ways in which testing and restudying increased final recall. Testing tends to strengthen correctly recalled items, which tend to be made with high confidence, and therefore testing helps to maintain the correspondence between initial confidence ratings and final recall. In contrast, restudying is more likely to benefit low-confident correct items and initially incorrect items, which are typically given low-confident ratings. Because restudying enhances performance of low-confident items, it decreases the correspondence between confidence and performance.

[10] Gamma correlations between confidence ratings and final test performance could not be calculated for seventeen participants who did not recall any items on the final test.

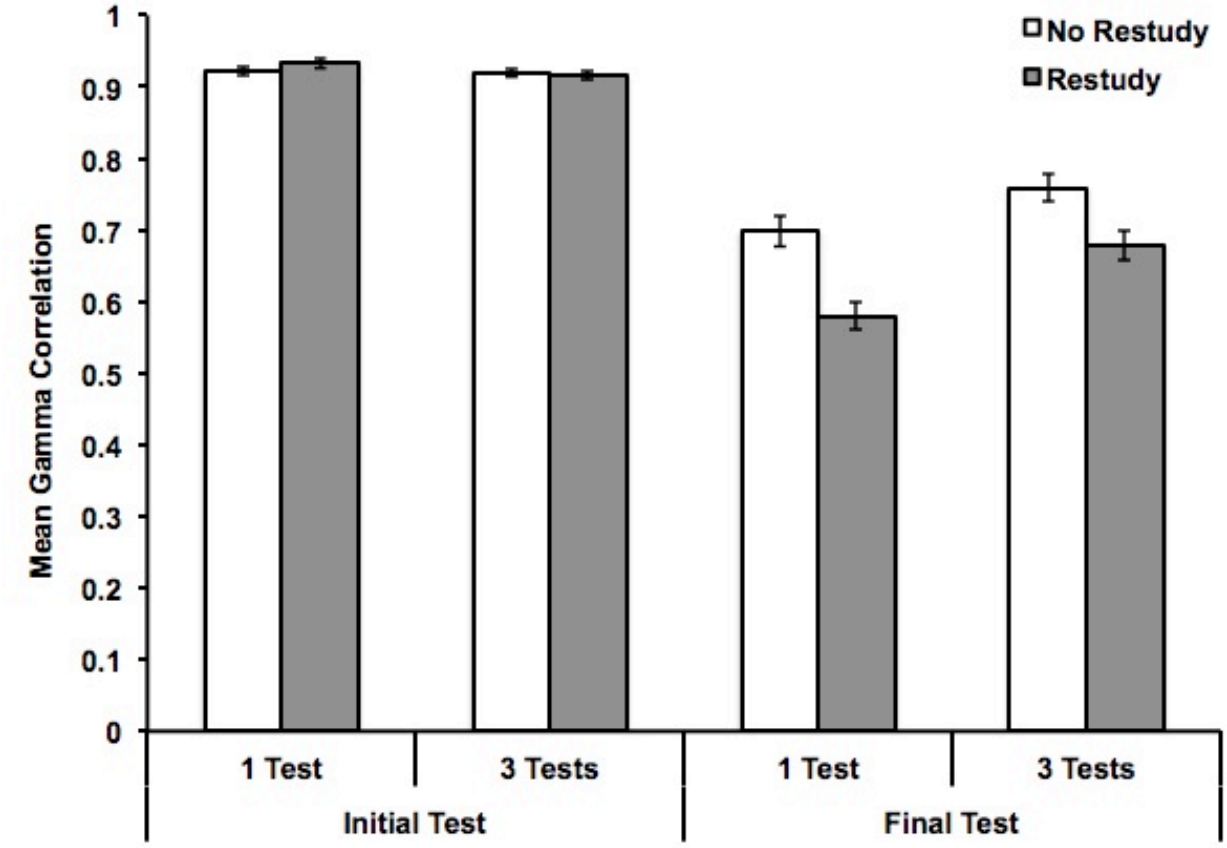

Figure 23. Mean gamma correlation between initial confidence ratings and initial test recall (left panel) and final test recall (right panel) in Experiment 2 as a function of test and restudy conditions. Error bars represent standard errors of the mean.

Post-experimental questionnaire

After both sessions, subjects were asked how interested they were in the study, how difficult they thought it was, and how much effort they put into it (see Table 1). Subjects answered these questions using a 5-point scale, with 1 meaning very little interest, difficulty, or effort and 5 meaning a great deal of interest, difficulty or effort. For each of these dependent variables, a 2 (test: 1, 3) X 2 (restudy, no restudy) X 2 (session: 1, 2) mixed ANOVA was conducted to determine if ratings differed across conditions and/or across sessions[11].

First, ratings of interest in the experiment were analyzed. Subjects in the 1 and 3 test conditions rated interest in the experiment differently across the two sessions as indicated by a test condition by session interaction, $F(1, 613) = 22.30$, $p < .001$, $\eta_p^2 = .04$. In Session 1, subjects who took one test found the experiment more interesting than those who took three tests (M = 3.93 vs. 3.63), $t(617) = 3.42$, $p = .001$, $d = .28$, indicating that taking three successive tests is less interesting than taking only one test. This difference disappeared by Session 2. In this session, subjects in both conditions rated the experiment as equally interesting (M = 3.77 vs. 3.82), $t < 1$, indicating that the effect of taking two additional tests during Session 1 did not carry over to Session 2. No other effects were significant for interest ratings.

Ratings of difficulty differed across sessions, $F(1, 613) = 52.30$, $p < .001$, $\eta_p^2 = .08$. Subjects tended to rate Session 2 (M = 4.23) as more difficult than Session 1 (M = 3.93). However, this increase was not consistent across test conditions, $F(1, 613) = 22.91$, $p < .001$, $\eta_p^2 = .04$. During Session 1, subjects in the 3 Test condition rated the experiment as more difficult than subjects in the 1 Test condition (M = 4.03 vs. 3.84), $t(617) = 2.61$, $p = .01$, $d = .21$. This

[11] Due to a technical error, post-experimental questionnaire data is incomplete for seven participants. For each participant, ratings are missing for interest, difficulty, or effort ratings for either Session 1 or 2.

suggests that in addition to making the experiment seem less interesting, taking two additional tests also made the experiment seem more difficult. However, during Session 2, this pattern reversed. Subjects in the 3 Test condition now found the experiment less difficult than subjects in the 1 Test condition (M = 4.13 vs. 4.33), $t(617) = 2.80$, $p = .01$, $d = .23$. This change was likely due to the effect of testing on final recall. Subjects in the 3 Test condition were able to recall more item on the final test than subjects in the 1 Test condition (see Figure 14), which likely made the experiment seem less difficult. No other effects were significant for difficulty ratings.

Effort ratings also varied across sessions. Participants reported using more effort during Session 1 (M = 4.65) than during Session 2 (M = 4.33), $F(1, 616) = 67.38$, $p < .001$, $\eta_p^2 = .10$. This difference may have been due to the difference in length and task requirements between the two sessions. Session 1 involved at minimum one study and one test phase, whereas Session 2 only involved a final test. No other effects were significant for the effort ratings.

Discussion

The primary aim of this experiment was to determine if tests potentiate subsequent learning of correctly recalled items. The results do not support this hypothesis. Although restudying after correct recall enhanced retention, the number of prior tests did not modify the effect of restudying. Further, tests did not modify the effect of restudying across any level of confidence. The effect of restudying was more pronounced for low-confident correct items and diminished as confidence increased, but this relationship was the same in both test conditions. Although this null finding is not direct evidence against the enhanced retention component of test-potentiated learning, it suggests that tests may not potentiate the subsequent encoding of correctly recalled items.

This is the first study to directly examine the enhanced retention component of test-potentiated learning. Although this study does not provide support for this effect, one null finding does not negate the possibility that tests may potentiate correctly retrieved items in some circumstances. For instance, in this study the effect of taking three prior tests was compared to the effect of taking one prior test. This may not be the optimal comparison for finding enhanced retention. The effect of correctly retrieving items one time may have a potentiating effect on subsequent study and correctly retrieving those same items two additional times may not substantially increase the impact of that effect. Unfortunately, the paradigm employed in Experiment 2 could not be used to compare a one test condition to a zero test condition because in a zero test condition, conditional probability could not be used to analyze the data given that no initial test would be taken.

A related possibility is that the measure used to gauge retention was not sensitive enough to detect an enhanced retention effect. In the current paradigm, finding evidence for enhanced retention depended on finding a significant ordinal interaction, which requires more power to find than a disordinal interaction (Bobko, 1986; Strube & Bobko, 1989). Given that confidence was highly negatively skewed for initially correct items, there may not have been enough power to find an enhanced retention effect. An enhanced retention effect may be found using a paradigm and/or material that generated more low-confident correct items. For example, multiple-choice tests using general knowledge questions (as in Butler et al., 2008) may produce more low-confident correct items, which may in turn allow an enhanced retention effect to be found.

Alternatively, a more sensitive final test may reveal an enhanced retention effect. For instance, a recognition test could allow subjects to use partial knowledge to determine the correct

target (Brown & McNeill, 1966), which may allow the enhanced retention effect to emerge. Subjects in the three tests condition may have retained some incomplete target information that those in the one test condition did not. A recognition test may be able to capture this additional partial information. Alternatively, a final test in which the subject is asked to recall the cue rather than the target may be more sensitive to an enhanced retention effect. This type of reverse cued-recall test has been shown to be more sensitive to some effects (e.g., Hays et al., 2012) and could, like a recognition test, reveal additional information that subjects in the three test condition may have obtained on the restudy trial.

Finally, this experiment only looked for enhanced retention in paired-associate learning with cued recall tests. A different kind of initial test could have different effects on initially correct items. For example, correctly retrieving an item on a free recall test could have a greater potentiating effect on subsequent restudy than correctly retrieving an item on a cued recall test. In general, retrieving an item with fewer cues requires more effort than retrieving the same item with more cues (e.g., Carpenter & Delosh, 2006). More effortful retrieval has been shown to have a greater direct effect on retention (Bjork & Bjork, 1992), and therefore more effortful retrieval may also have a greater indirect effect on retention. For this reason, more effortful retrieval may lead to more test-potentiation of initially correct items.

General Discussion

The primary aim of the present dissertation was to enhance understanding of test-potentiated learning by examining two ways in which tests may potentiate learning: by enhancing encoding and enhancing retention. Experiment 1 focused on the first component and found evidence corroborating previous findings that tests enhance subsequent encoding of initially incorrect items. Further, Experiment 1 extended previous findings by determining that

making unsuccessful retrieval attempts during initial tests is what causes subsequent encoding to be enhanced (but see Footnote 7). Experiment 2 was the first experiment to specifically examine the enhanced retention component of test-potentiated learning. Specifically, this experiment investigated the hypothesis that prior retrieval practice may modify the effect of restudying correct items on their later retention. The results failed to support this hypothesis, suggesting that prior tests may not enhance subsequent encoding of initially correct items. Together, these two experiments expand our understanding of test-potentiated learning and provide a stronger foundation upon which test-potentiated learning theories can be formed.

Possible Mechanisms Underlying Test-Potentiated Learning

The results from these experiments indicate that test-potentiated learning is driven by unsuccessful retrieval. This suggests that test-potentiated learning depends on processes that occur during an unsuccessful retrieval attempt. Several different theories have suggested some processes that could occur during an unsuccessful retrieval attempt and could, in turn, enhance subsequent encoding. For example, retrieval searches that fail to result in correct recall could nonetheless lead to activation of the correct target through spreading activation (Collins & Loftus, 1975; Grimalidi & Karpicke, 2012; Kornell et al., 2009). Alternatively, unsuccessful retrieval could suppress the association between the cue and the correct target, which could paradoxically enhance subsequent encoding (Hays et al., 2012; Kornell et al., 2009). Another possibility is that incorrectly retrieved items could act as a mediator between the cue and the correct target and could thus serve as an additional retrieval cue for the correct target on a later test (Kornell et al., 2009). A different theory suggests that unsuccessful retrieval attempts enhance metacognitive knowledge, which can then be used during subsequent restudy to enhance learning (Arnold & McDermott, in preparation; Lachman & Laughery, 1968; LaPorte & Voss, 1974; Royer, 1973).

Finally, unsuccessful retrieval attempts could subsequently enhance study-phase retrieval or remindings during restudy, which could in turn enhance learning (Nelson et al., under review).

Given the results from Experiment 1 along with results from previous experiments, only two of these theories (enhanced metacognition and enhanced remindings) remain viable candidates. The first three theories (spreading activation, suppression of cue-target association, and incorrect guess as mediator), which were all developed to explain generate-potentiated learning, cannot explain many test-potentiated learning results. In particular, spreading activation cannot account for why test-potentiated learning can occur when there is a delay of several minutes and/or several intervening items between the retrieval attempt and the restudy opportunity, and all three theories cannot explain how tests can potentiate learning of unrelated pairs. In contrast, both the enhanced metacognition theory and the enhanced remindings theory can account for these kinds of data. Future research is needed to determine which (if either) of these processes underlie test-potentiated learning.

These theories are not mutually exclusive. Multiple processes may work together to create test-potentiated learning. Further, different processes may create test-potentiated learning under different circumstances (e.g., cued recall tests and free recall tests may enhance subsequent learning in different ways). The present dissertation exclusively focused on the potentiating effect of cued recall tests and found that unsuccessful retrieval drove the effect, but the same might not be true for free recall tests. On a free recall test, there are not distinct trials devoted to each item, and therefore failing to retrieve an item on a free recall test does not necessarily indicate that an unsuccessful retrieval attempt devoted to that item was made. For an item that is not retrieved on a free recall test, no specific cue is provided to prompt the subject to make a retrieval attempt. Yet despite this less constrained retrieval context, free recall tests have been

shown to enhance subsequent learning (Arnold & McDermott, 2013). For this reason, free recall tests may enhance learning through successful retrieval, such as by enhancing the organization of retrieved items, rather than, or in addition to, through unsuccessful retrieval attempts.

Another way in which cued recall and free recall tests could have differential effects is through potentiation of initially correct items. Although Experiment 2 found no evidence of the enhanced retention component of test-potentiated learning with cued recall tests, if different processes underlie cued recall and free recall test-potentiated learning, free recall tests could potentiate learning of initially correct items. For instance, if free recall tests potentiate learning by enhancing organization of recalled items as suggested by Arnold and McDermott (2013), both initially incorrect and initially correct items could benefit from this improved structure during subsequent study. Future studies are needed to determine if different processes drive potentiation in free recall and cued recall tests, and if the processes are different, future studies are needed to determine whether or not they affect subsequent study differently.

The Relation between the Testing Effect with Feedback and Test-Potentiated Learning

Many studies examining the beneficial effects of retrieval practice employ feedback after retrieval attempts (e.g., Butler et al., 2008; Kang, McDermott, & Roediger, 2007; Pashler et al., 2005). When feedback involves presenting the correct answer or target item, it can be viewed as a restudy opportunity. For this reason, test-potentiated learning is likely present in these experiments. Anytime an unsuccessful retrieval attempt is made prior to a restudy opportunity, test-potentiated learning may occur. However, as discussed earlier, control conditions and/or particular analyses such as conditional probability analyses are needed to distinguish the indirect potentiating effects of prior testing from the direct effects of prior testing. Many experiments that

use feedback are not designed in such a way that these types of effects can be disambiguated, and therefore even though test-potentiated learning may be present, it cannot always be measured.

Experiments that involve testing with feedback are often not designed to measure test-potentiated learning because they typically ask a different kind of question (e.g., for reviews see Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kulhavy, 1977). Rather than asking how does taking a prior test modify the effect of a subsequent restudy opportunity, they ask how does providing a subsequent restudy opportunity modify the effect of taking a prior test. This point is illustrated in Figure 24. Studies that ask the latter question typically manipulate the restudy phase in some way to determine how that manipulation modifies the effect of the initial test (top panel of Figure 24). For instance, Butler and Roediger (2008) had subjects take a multiple-choice test with either immediate, delayed, or no feedback. They found that both immediate and delayed feedback enhanced final test performance relative to the no feedback condition. Specifically, feedback enhanced the proportion of correct responses and reduced the proportion of intrusions (i.e., incorrect lures on the initial multiple choice test) on a delayed final cued recall test. The authors concluded that "feedback on a multiple-choice test . . . enhance[s] the testing effect for items answered correctly and reduce[s] or eliminate[s] negative effects on items answered incorrectly" (p. 611). In other words, restudying modifies the effect of the initial test.

This approach is contrasted with studies examining test-potentiated learning. These studies typically manipulate the initial test phase to determine how that manipulation modifies the effect of the subsequent restudy phase (bottom panel of Figure 24). For instance, in Experiment 1 of the present experiments, the amount of unsuccessful and successful retrieval subjects engaged in prior to restudying was manipulated to determine how these types of retrieval attempts affect

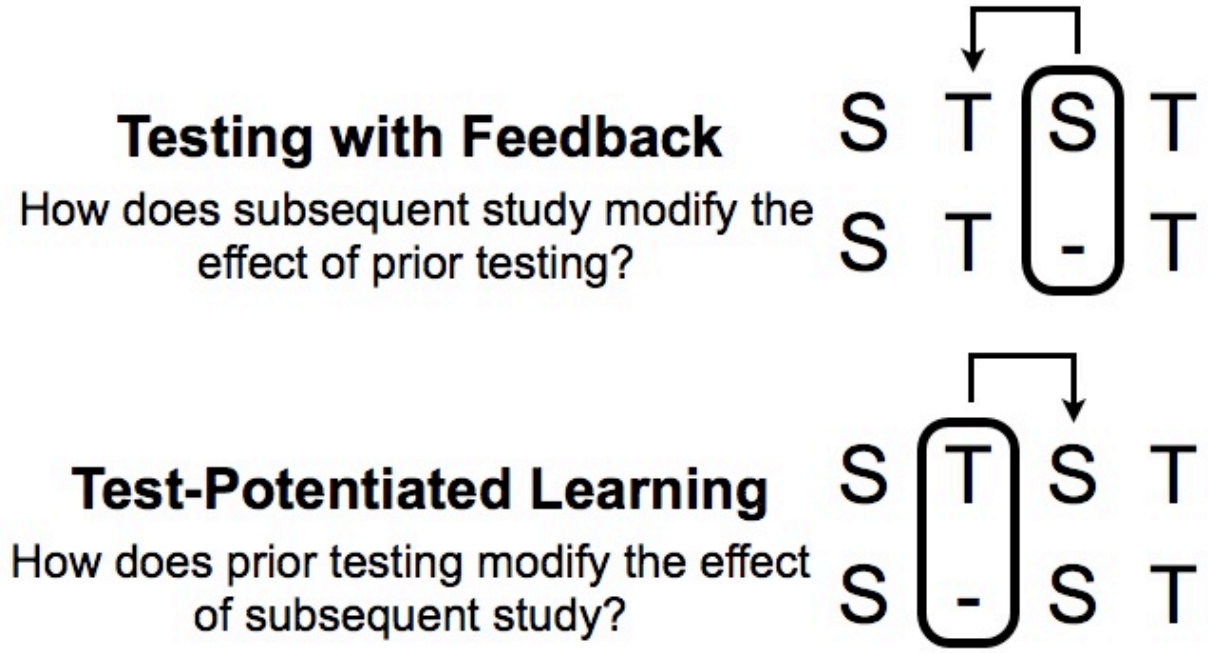

Figure 24. The difference between studies examining feedback (top panel) and studies examining test-potentiated learning (bottom panel). A box surrounds the phase that is manipulated. An arrow points from the phase that is manipulated to the phase of primary interest. S = study period, T = test period, - = blank period.

learning during the subsequent restudy phase. The results indicated that unsuccessful retrieval enhanced subsequent learning. In other words, prior retrieval attempts modified the effect of subsequent restudy.

Studies that use feedback when examining the beneficial effects of prior retrieval practice may be enhanced by considering how test-potentiated learning may contribute to recall on the final test. For example, Butler and Roediger (2008) concluded that feedback enhanced the testing effect. However, there is an alternative interpretation; feedback may not so much have enhanced the testing effect as provided a situation in which the testing effect and test-potentiated learning could both contribute to later performance. That is, together the testing effect and the test-potentiated learning effect have a greater enhancing effect on memory than the testing effect alone.

Two experiments by Kang et al. (2007) further illustrate this point. Although test-potentiated learning was not measured in these experiments, considering this effect may help explain the results. In their first experiment, after reading an article, some subjects either took an initial cued recall test without feedback or read statements summarizing the article. Three days later, subjects took a final test. The typical advantage of testing over rereading was not found. Performance on the final test was either equivalent in the reread and cued recall test conditions or slightly better in the reread condition than in the cued recall test condition depending on the format of the final test. The authors pointed to low initial test performance as a possible reason for these results. Only about half of the items were correctly recalled on the initial cued recall test, and therefore subjects were not re-exposed to the other half of the items. To test this idea, in their second experiment, they repeated the same procedure but provided feedback after the initial test. Initially tested items were now better remembered on the final test than items that were only

reread. When feedback was provided, an initial cued recall test benefited items even when those items were not recalled on this initial test. The results suggest that making an initial retrieval attempt before rereading enhanced later performance relative to rereading without first making an initial retrieval attempt, suggesting that test-potentiated learning may have taken place. Although the authors concluded that feedback modified the effect of testing, the results could be interpreted in the converse direction. That is, testing modified the effect of feedback. Subjects were able to learn more from the feedback (or reread statements) when they followed an initial retrieval attempt.

Educational Applications of Test-Potentiated Learning

Experiments measuring test-potentiated learning have so far been limited to contrived laboratory settings. There are many differences between this type of setting and a real educational context. For instance, the materials used in these studies have tended to be simple. Subjects typically learn lists of words, pictures, or paired items (e.g., Arnold & McDermott, 2012; 2013; Izawa, 1966, 1968, 1971; Karpicke & Roediger, 2007b; Tulving, 1967). Further, the consequences of learning or not learning this simple material are minimal or non-existent, and the retention intervals over which this material needs to be remembered are fairly short. Even when a final test is delayed by one week as in the case of Experiment 2, the material has to be retained for a much shorter period of time than would be expected in an educational context. Because of these (and other) differences, any recommendations that can be made for how or if test-potentiated learning should be incorporated into educational settings are at this time mainly conjecture and are very preliminary.

Even though recommendations cannot be made with any certainty, the results from the experiments presented here and other test-potentiated learning experiments do hint at possible

educational implications. Testing may benefit the learning process early on and not just after a set of material has been mastered. For instance, questions asked throughout a lecture that encourage students to retrieve already presented material could aid in learning. After these retrieval attempts are made, learning will likely be enhanced if feedback is provided. Allowing students an opportunity to restudy what they just attempted to retrieve may allow them to benefit from test-potentiated learning. In this way, students can maximize the benefit of the retrieval attempt.

Future studies done in the classroom could test this hypothesis by manipulating both retrieval practice and feedback (i.e., restudy). For instance, an experiment could examine the effect of asking frequent questions (with required responses by every student to encourage retrieval attempts) throughout a lecture followed by feedback to the effect of asking the questions without feedback, providing the feedback without the questions, and neither asking the questions nor providing feedback. This 2 (questions, no questions) X 2 (feedback, no feedback) factorial design is similar to the one used in Experiment 2 (see Figure 12). If prior retrieval attempts enhance learning during feedback, the difference in final recall between the feedback and no feedback condition should be larger when questions were asked throughout the lecture. This result would suggest that the retrieval attempts made in response to the questions potentiated learning during feedback.

One way in which testing prior to mastery of a topic is already commonly incorporated into many educational contexts is through pre-testing. Many educators employ pre-testing as a tool to aid in learning a new topic (Pashler, Bain et al., 2007). However, research on the utility of pre-tests has been mixed (for reviews see Anderson & Biddle, 1975; Hamaker, 1986). Some researchers have found pre-tests have no benefit to learning or only benefit learners to the extent

that they direct attention to particular items (e.g., McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; Rothkopf, 1966). Many of these studies have shown that when pre-tests benefit final performance for tested items, they tend to hurt performance on non-tested items. This finding is likely due to subjects directing their attention to tested items and ignoring, or decreasing attention, to non-tested items. In contrast, other studies have shown that pre-testing can enhance learning for tested items without hurting performance on non-tested items, suggesting that pre-testing may have an effect beyond directing attention (e.g., Pressely, Tanenbaum, McDaniel, & Wood, 1990; Richland, Kornell, & Kao, 2009).

Studies on test-potentiated learning along with those on generate-potentiated learning may be able to provide some insight into why pre-testing seems to help in some cases and have no or minimal benefit in other cases. When pre-tests are used before students have had any exposure to a topic, they have no episodic memory upon which to draw from when making a retrieval attempt. In these cases, the paradigm may be more akin to a generate-potentiated learning paradigm than to a test-potentiated learning paradigm. If this is the case, then the boundary conditions of generate-potentiated learning may prevent students from benefiting from the pre-test. In particular, if the student does not have a pre-existing association between the target answer and the question (i.e., if the target and question are unrelated from the student's perspective) or if the answer is provided after a delay, pre-testing may not enhance learning (see Grimaldi & Karpicke, 2012; Huesler & Metcalfe, 2012). In contrast, if the pre-test is given after subjects have some minimal exposure to the material (e.g., after they have read a textbook chapter), they may be able to capitalize on test-potentiated learning. If so, different boundary conditions may put different constraints on the benefit of the pre-test. For instance, the pre-test may enhance learning even if the answer is provided after a short delay.

Future Directions: Unanswered Questions about Test-Potentiated Learning

Test-potentiated learning has been an under-researched effect, and many important questions remain unanswered. Some of these unanswered questions pertain to how unsuccessful retrieval can enhance subsequent learning. Some possible underlying processes have been described in earlier sections of this dissertation. Determining which processe(s) drive this effect could provide valuable insight into not only the consequences and limitations of test-potentiated learning but also into the overall understanding of human memory. The relationship between encoding and retrieval plays a vital role in our ability to remember. Test-potentiated learning is one of many examples of how these two important processes interact with one another, and understanding the nature of this particular interaction could add to the general understanding of how these processes interact.

Other unanswered questions pertain to limitations of test-potentiated learning. Answers to these questions have important implications for the applied value of the effect. For instance, the possible educational importance of test-potentiated learning depends on the long-term viability of the effect. That is, is the material that is learned through test-potentiated learning remembered over a long retention interval? As was pointed out in the previous section, retention intervals tend to be fairly short in test-potentiated learning experiments, and often the final test is given within the same session as the initial learning phase (e.g., Arnold & McDermott, 2012; 2013; Donaldson, 1971; Izawa, 1966; 1971; Karpicke & Roediger, 2007b; Lachman & Laughery, 1968). Because of this, the long-term benefit of test-potentiated learning is unknown.

Experiment 2 of the present dissertation is one of the few test-potentiated learning studies with a delayed final test. This experiment suggested that the effects of the enhanced encoding component of test-potentiated learning might last over at least a week, although these results are

far from conclusive. As can be seen in Figures 20, 21, and 22, for subjects who restudied, a numerically larger proportion of initially incorrect items were recalled on the final test in the 3 Test condition than in the 1 Test condition, suggesting that the additional tests potentiated subsequent learning of these items and that the effects of this potentiated learning was still apparent a week later. However, floor effects present in the data hinder interpretation of these results. Further, different statistical analyses indicated different significant effects. When confidence was not taken into account, an ANOVA indicated a marginally significant test condition by restudy condition interaction, suggesting that the differences between the 1 and 3 Tests conditions for subjects who restudied may have been meaningful. In contrast, when confidence was included as a variable, the logistic hierarchical linear regression analysis gave no indication of an interaction between these variables. Further research in which floor effects are not present is needed to determine the lasting effects of enhanced encoding.

A related unanswered question pertains to the delay between the initial test(s) and the restudy opportunity. How long of a delay can there be between a test and subsequent restudy for there to still be a potentiating effect? Research on generate-potentiated learning suggests that after a failed generation attempt, the restudy opportunity must happen immediately for the test to affect learning (Grimaldi & Karpicke, 2012; but see Hays et al., 2012). In contrast, test-potentiated learning experiments have demonstrated an effect of a failed retrieval attempt when the restudy opportunity has been delayed by several minutes (e.g., Arnold & McDermott, 2012; 2013; Izawa, 1966; 1971). However, a delay of several minutes is still a fairly short delay compared to how long after a test a restudy opportunity might occur in educational contexts. For instance, a delay of several days may pass after students take a formal test before a teacher may provide feedback on that test. The merits of providing feedback immediately as compared to

providing it after a delay has been debated in the testing effect literature for some time (for a review see Kulik & Kulik, 1988). Some researchers have found that delayed feedback is more beneficial than immediate feedback (e.g., Butler et al., 2007), which suggests that test-potentiated learning may last over such a delay. Determining if tests can enhance learning on a delayed restudy opportunity could add important information to the debate on the merits of immediate versus delayed feedback.

In addition to issues relating to delay, many other questions about test-potentiated learning remain. For instance, as has been previously noted, most test-potentiated learning experiments use simple stimuli (e.g., Arnold & McDermott, 2012; 2013; Izawa, 1966; 1971). Can tests potentiate learning of more complicated material? For instance, could testing enhance learning of prose passages or complicated concepts (e.g., scientific theories and principles)? As with the questions about test-potentiated learning over delays, answers to these questions could have significant impact on the practicality of capitalizing on this effect in educational contexts. If tests can only potentiate learning of simple material, test-potentiated learning would have little value in most applied settings.

Another important question is how tests may or may not affect learning of untested material. Some pre-test studies show that pre-testing may actually hurt learning of untested material (e.g., Anderson & Biddle, 1975), whereas others indicate they have no effect on untested material (e.g., Richland et al., 2009). Most studies, however, do not indicate that pre-tests can benefit untested material. The same may be true for test-potentiation of untested material. Preliminary evidence indicates that tests do not potentiate items that are not directly tested (Arnold, Nelson, & McDermott, in preparation). If this finding is confirmed, the utility of test-potentiated learning in educational contexts may be limited. That is not say, though, that

test-potentiated learning would not still have value in classrooms. Teachers could focus testing on the most important points and/or on the most difficult material.

Conclusion

The effect of testing, or retrieval practice, has been the focus of numerous experiments in recent years (e.g., Roediger & Karpicke, 2006b). Yet, despite this focus, test-potentiated learning remains an underappreciated effect. One goal of the present research was to bring attention to this important effect and to encourage researchers to recognize when test-potentiated learning may be present in their own paradigms. For example, test-potentiated learning may affect results in many retrieval practice experiments that include feedback. Just as feedback may modify the effect of testing, testing may also modify the effect of feedback. The latter effect may be just as, if not more, important to learning and memory.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs, 11,* 159–177.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1063–1087.

Anderson, M. C., Bjork, E. L., & Bjork, R. A. (2000). Retrieval-induced forgetting: Evidence for a recall-specific mechanism. *Psychonomic Bulletin & Review, 7*, 522–530.

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. Bower (Ed.), *Annual review of psychology* (Vol. 18, pp. 129-164). Palo Alto, CA: Annual Reviews.

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1972). Conditions under which feedback facilitates learning from programmed lessons. *Journal of Educational Psychology*, *63*, 186–188.

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8*, 63–470.

Arnold, K. M., & McDermott, K. B. (2012). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Advanced online publication.

Arnold, K. M., & McDermott, K. B. (2013). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review*. Advanced online publication.

Arnold, K. M., & McDermott, K. B. (in preparation). Why do tests enhance subsequent learning?

Arnold, K. M., Nelson, S. M., & McDermott, K. B. (in preparation). Test-potentiated learning is specific to tested items.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213-238.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: University Press.

Battig, W. F., Allen, M., & Jensen, A. R. (1965). Priority of free recall of newly learned items. *Journal of Verbal Learning and Verbal Behavior, 4*, 175–179.

Bäuml, K.-H. T., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language, 68*, 39–53.

Birnbaum, I. M., & Eichner, J. T. (1971). Study versus test trials and long-term retention in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *10*, 516–521.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. Kosslyn & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

Bobko, P. (1986). A solution to some dilemmas when testing hypotheses about ordinal interactions. *Journal of Applied Psychology, 71*, 323–326.

Bregman, A. S., & Wiener, J. R. (1970). Effects of test trials in paired-associate and free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *9*, 689–698.

Brown, R., & McNeill, D. (1966). The “tip of the tongue” phenomenon. *Journal of Verbal Learning and Verbal Behavior, 5*, 325–337.

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bulletin & Review, 18*, 1238–1244.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*, 273–281.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 918–928.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*, 604–616.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1491–1494.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition Learning, 1*, 69–84.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*, 268–276.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*, 633–642.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354–380.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*, 407–428.

Crowder, R. G. (1976). *Principles of learning and memory.* Hillsdale, NJ: Erlbaum.

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review, 1*, 309–330.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Handbook of perception and cognition: Memory* (pp. 317–344). San Diego, CA: Academic Press.

Donaldson, W. (1971). Output effects in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 577–585.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology*. (H. A. Ruger & C. E. Bussenius, Trans.). New York: Dover. (Original work published 1885).

Erdelyi, M. H., & Becker, J. (1974). Hypermnesia for pictures: Incremental memory for pictures but not words in multiple recall trials. *Cognitive Psychology, 6*, 159–171.

Estes, W. K. (1964). All-or-none processes in learning and retention. *American Psychologist*, *19*, 16–25.

Evans, J. J., Wilson, B. A., Schuwi, U., Andrade, J., Baddeley, A., Bruna, O., Canava, T., Della Sala, S., Green, R., Laaksonen, R., Lorenzi, L., & Taussik, I. (2000). A comparison of "errorless" and "trial-and-error" learning methods for teaching individuals with acquired memory deficits. *Neuropsychological Rehabilitation, 10*, 67–101.

Gardiner, J. M., & Klee, H. (1976). Memory for remembered events: An assessment of output monitoring in free recall. *Journal of Verbal Learning and Verbal Behavior, 15*, 227–233.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, *5*, 351–360.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior, 15*, 1–16.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 2*, 119–136.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 371–377.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40*, 505–513.

Guthrie, E. R. (1952). *The psychology of learning*. New York: Harper.

Guthrie, J. T. (1971). Feedback and sentence learning. *Journal of Verbal Learning and Verbal Behavior, 10*, 23–28.

Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research, 56*, 212-242.

Hays, M. J., Kornell, N., & Bjork, R. A. (2012). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Advanced online publication.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.). *Theories in cognitive psychology: The Loyola symposium* (p. 77–99). Hillsdale, New Jersey: Lawrence Eribaum Assoc.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior,* 10, 562–567.

Huelser B. J., & Metcalf, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition, 4*, 514–527.

Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports*, *18*, 879–919.

Izawa, C. (1967a). Function of test trials in paired-associate learning. *Journal of Experimental Psychology*, *75*, 194–209.

Izawa, C. (1967b). Mixed- versus unmixed-list designs in paired-associate learning. *Psychological Reports, 20*, 1191–1200.

Izawa, C. (1968). Effects of reinforcement, neutral and test trials upon paired-associate acquisition and retention. *Psychological Reports*, *23*, 947–959.

Izawa, C. (1969a). Comparison of reinforcement and test trials in paired-associate learning. *Journal of Experimental Psychology, 81*, 600–603.

Izawa, C. (1969b). Long sequences of successive tests in paired-associate learning acquisition. *Proceedings of the 77th Annual Convention of the American Psychological Association, 4*, 57–58.

Izawa, C. (1970a). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology*, *83*, 340–344.

Izawa, C. (1970b) Reinforcement-test-blank acquisition programming under the unmixed list design in paired-associate learning. *Psychonomics Science, 19*, 340–344.

Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology*, *8*, 200–224.

Kang, S. H., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*, 528-558.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*, 469–486.

Karpicke, J. D., & Roediger, H. L. (2007a). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 704–719.

Karpicke, J. D., & Roediger, H. L. (2007b). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151–162.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968.

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*, *62*, 227–239.

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology, 93*, 329–343.

Knecht, S. Breitenstein, C., Bushuven, S., Wailke, S., Kamping, S., Flöel, A., Zwitserlood, P., & Ringelstein, E. B. (2004). Levodopa: Faster and better word learning in normal humans. *Annals of Neurology, 56*, 20–26.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*, 219–224.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition; Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 989–998.

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*, 493–501.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*, 211–232.

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology*, *63*, 505–512.

Kulhavy, R. W., Yekovich, F. R., & Dyer, J. W. (1976). Feedback and response confidence. *Journal of Educational Psychology, 68*, 522–528.

Kulik, J. A., & Kulik, C. C., (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*, 79–97.

Lachman, R., & Laughery, K. R. (1968). Is a test trial a training trial in free recall learning?. *Journal of Experimental Psychology*, *76*, 40–50.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

LaPorte, R., & Voss, J. F. (1974). Paired-associate acquisition as a function of number of initial nontest trials. *Journal of Experimental Psychology*, *103*, 117–123.

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior, 8*, 828–835.

Mandler, G. (1967). Organization and memory. In K. W. Spence and J. T. Spence (Eds.), *The Psychology of Learning and Motivation* (Vol. 1, pp. 327–372), New York: Academic Press.

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition, 39*, 462–476.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*, 399-414.

McDermott, K. B., Arnold, K. M., & Nelson, S. M. (in press). The testing effect. In T. Perfect & S. Lindsay (Eds.), *Sage Handbook of Applied Memory*.

McDermott, K. B., Szpunar, K. K., & Christ, S. E. (2009). Laboratory-based and autobiographical retrieval tasks differ substantially in their neural substrates. *Neuropsychologia, 47*, 2290–2298.

McNamara, T. P. (1992). Theories of Priming: I. Associative Distance and Lag. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 1173-1190.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior, 9*, 596–606.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*, 174–179.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63*, 81–97.

Murdock, B. B. (1961). The retention of individual items. *Journal of Experimental Psychology, 62*, 618–625.

Neely, J. H., O'Connor, P. A., & Calabrese, G. (2010). Fast trial pacing in a lexical decision task reveals a decay of automatic semantic activation. *Acta Psychologica, 133*, 127–136.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods*, *36*, 402–407.

Nelson, S. M., Arnold, K. M., Gilmore, A. W., & McDermott, K. B. (under review). A neural signature of test-potentiated learning in parietal cortex.

Nelson, S. M., Cohen, A. L., Power, J. D., Wig, G. S., Miezin, F. M., Wheeler, M. E., Velanova, K., Donaldson, D. I., Phillips, J. S., Schlaggar, B. L., & Petersen, S. E. (2010). A parcellation scheme for human left lateral parietal cortex. *Neuron, 67*, 156–170.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*, 267–270.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 676–686.

Nunes, L. D., & Weinstein, Y. (2012). Testing improves true recall and protects against the build-up of proactive interference without increasing false recall. *Memory, 20*, 138–154.

Packard, M. G., & White, N. M. (1991). Dissociation of hippocampus and caudate nucleus memory systems by posttraining intracerebral injection of dopamine agonists. *Behavioral neuroscience*, *105*, 295–306.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 3–8.

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing Instruction and Study to Improve Student Learning* (NCER 2007-2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ncer.ed.gov.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, *14*, 187–193.

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K.-H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 287–297.

Payne, D. G. (1987). Hypermnesia and reminiscence in recall: A historical and empirical review. *Psychological Bulletin, 101*, 5–27.

Postman, L., & Keppel, G. (1977). Conditions of cumulative proactive inhibition. *Journal of Experimental Psychology: General*, *106*, 376–403.

Postman, L., & Schwartz, M. (1964). Studies of learning to learn. I. Transfer as a function of method of practice and class of verbal materials. *Journal of Verbal Learning and Verbal Behavior*, *3*, 37–49.

Pressley, Tanenbaum, R., McDaniel, M. A., Wood, E. (1990). What happens when university students try to answer prequestions that accompany textbook material? *Contemporary Educational Psychology, 15*, 27-35.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*, 335.

Pyc, M. A., & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 737–746.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15*, 243-257.

Rock, I. (1957). The role of repetition in associative learning. *The American Journal of Psychology*, 186–193.

Roediger, H. L., & Arnold, K. M. (2012). The one-trial learning controversy and its aftermath: Remembering Rock (1957). *The American Journal of Psychology*, *125*, 127–143.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, *15*, 20–27.

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21***,** 803–814.

Roediger, H. L., & Smith, M. A. (2012). The "pure-study" learning curve: The learning curve without cumulative testing. *Memory & Cognition*, *40*, 989–1002.

Rosner, S. R. (1970). The effects of presentation and recall trials on organization in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior, 9*, 69–74.

Rothkopf, E. Z. (1966). Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal, 3*, 241-249.

Royer, J. M. (1973). Memory effects for test-like-events during acquisition of foreign language vocabulary. *Psychological Reports, 32*, 195–198.

Shaughnessy, J. J., & Zechmeister, E. B. (1992). Memory-monitoring accuracy as influenced by the distribution of retrieval practice. *Bulletin of the Psychonomic Society, 30*, 125–128.

Skinner, B. F. (1958). Teaching machines: From the experimental study of learning come devices which arrange optimal conditions for self-instruction. *Science, 128*, 969–977.

Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior*, *22*, 153–163.

Smith, A. D. (1971). Output interference and organized recall from long-term memory. *Journal of Verbal Learning and Verbal Behavior, 10*, 400–408.

Soraci, S. A., Carlin, M. T., Chechile, R. A., Franks, J. J., Wills, T., & Watanabe, T. (1999). Encoding variability and cuing in generative processing. *Journal of Memory and Language*, *41*, 541–559.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641–656.

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2008). Accelerated relearning after retrieval-induced forgetting: The benefit of being forgotten. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 230–236.

Strube, M. J., & Bobko, P. (1989). Testing hypotheses about ordinal interactions: Simulations and further comments. *Journal of Applied Psychology, 74*, 247–252.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1392–1399.

Thiede, K. W., Anderson, M. C. M., Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66–73.

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior, 15*, 529–536.

Thomas, A. K., & McDaniel, M. A. (2007). Metacomprehension for educationally relevant materials: Dramatic effects of encoding-retrieval interactions. *Psychonomic Bulletin & Review*, *14*, 212–218.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 210–221.

Thorndike, E. L. (1912). The curve of work. *Psychological Review, 19*, 165-194.

Thorndike, E. L. (1914). Repetition versus recall in memorizing vocabularies. *Journal of Educational Psychology, 5*, 596–597.

Tulving, E. (1962). Subjective organization in free recall of "unrelated" words. *Psychological Review,* 69, 344–354.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*, 175–184.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving and W. Donaldson (Eds.) *Organization of memory* (pp. 381–402). New York: Academic Press.

Tulving, E. (1983). *Elements of episodic memory.* New York: Clarendon Press.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26*, 1–12.

Tulving, E., & Arbuckle, T. Y. (1963). Sources of intratrial interference in immediate recall of paired associates. *Journal of Verbal Learning and Verbal Behavior*, *1*, 321–334.

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*, 49–60

Wahlheim, C. N., & Jacoby, L. L. (2013). Remembering change: The critical role of recursive remindings in proactive effects of memory. *Memory & Cognition, 41*, 1–15.

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face-name learning. *Psychonomic Bulletin & Review, 18*, 518–523.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3*, 240–245.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review, 18*, 1140–1147.

Young, J. L. (1971). Reinforcement-test intervals in paired-associate learning. *Journal of Mathematical Psychology, 8*, 58–81.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*, 995–1008.

Appendix A

The 30 Indonesian-English word pairs used in Experiment 1.

Indonesian Word	English Word
kue	pastry
telur	egg
gendang	drum
emas	gold
sungai	river
susu	milk
bau	stink
ban	tire
satu	one
selatan	south
babi	pig
cepat	fast
sapi	cow
rok	skirt
seni	art
berkeliling	tour
bukit	hill
nasi	rice
mungil	cute
kiri	left
kunci	lock
ibu	mother
makanan	food
ombak	wave
roti	bread
danau	lake
panas	hot
keranjang	basket
angin	wind
buka	open

Appendix B

The list of stimuli used in Experiment 2. Groups of related words pairs are presented together. For each group of words pairs, the forward strength association between each cue and its paired target and the forward strength association between the cue of the second pair (Pair B) and the target of the first pair (Pair A) are presented.

		Forward Strength	
Cue	**Target**	**Paired Target**	**Related Target**
MOVIE	video	0.010	
TELEVISION	tube	0.014	0.021
IMPATIENCE	virtue	0.013	
HONESTY	love	0.014	0.027
INLET	shore	0.014	
BAY	river	0.013	0.027
GLORY	medal	0.011	
BRAVERY	valor	0.014	0.028
ENVY	wish	0.013	
BLESSING	gift	0.014	0.028
BUTTER	dairy	0.012	
CHEESE	protein	0.012	0.030
BACKGROUND	memory	0.013	
PICTURE	portrait	0.020	0.031
PANTS	slacks	0.011	
TROUSERS	shirt	0.019	0.032
DANCER	fame	0.013	
FORTUNE	happiness	0.014	0.042
KLEENEX	napkin	0.013	
HANDKERCHIEF	red	0.014	0.043
SHAPE	curve	0.011	
SLOPE	nose	0.013	0.046

FACT	myth	0.010	
LEGEND	unicorn	0.013	0.047
ORCHARD	peach	0.012	
PEAR	friend	0.020	0.048
NOODLES	gravy	0.013	
SAUCE	mushroom	0.014	0.050
CHORE	hobby	0.011	
CRAFTS	paint	0.013	0.052
DAGGER	blade	0.010	
KNIFE	pocket	0.013	0.122
JEEP	wheel	0.014	
BICYCLE	vehicle	0.014	0.122
NATION	flag	0.012	
PATRIOT	citizen	0.013	0.123
RHYTHM	poem	0.010	
PROSE	fiction	0.013	0.133
ADVICE	idea	0.010	
SUGGESTION	option	0.014	0.142
CITY	crowd	0.010	
PEOPLE	world	0.014	0.155
FOOD	mouth	0.011	
LIPS	teeth	0.014	0.185
PHOTO	color	0.010	
PRISM	pyramid	0.014	0.192
NERD	candy	0.013	
CHOCOLATE	fudge	0.014	0.196
GIRAFFE	trees	0.013	
LANDSCAPE	mountain	0.013	0.199
POUCH	wallet	0.010	

PURSE	leather	0.014	0.203
PINK	shoes	0.011	
PAIR	twin	0.020	0.208
SQUID	shrimp	0.013	
SEAFOOD	pasta	0.014	0.245
GROVE	bush	0.012	
HEDGE	lawn	0.020	0.309
KNIGHT	castle	0.013	
MOAT	pond	0.013	0.320